The Dissertation Committee for Hamidreza Arabshahi
certifies that this is the approved version of the following dissertation:

# Space-Time Hybridized Discontinuous Galerkin Methods for Shallow Water Equations

Committee:

Clinton Dawson, Supervisor

Ivo Babuška

Leszek Demkowicz

Patrick Heimbach

Alexis Vasseur

# Space-Time Hybridized Discontinuous Galerkin Methods for Shallow Water Equations

**by**

**Hamidreza Arabshahi, B.Sc.;M.Sc.;M.S.C.S.E.M.**

**DISSERTATION**

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

**DOCTOR OF PHILOSOPHY**

THE UNIVERSITY OF TEXAS AT AUSTIN

August 2016

*"We even made it to the stars but as the old saying went there was no old bearded god there, only science, only the Soviet Union"*

From the Soviet space program, 1965

# Acknowledgments

I would like to thank my advisor for his support during my study at ICES.

I had the opportunity to run the professor Babuška's forum for two years. This was enjoyable, gave me a chance to talk with and learn from him.

Thanks to Alexis Vasseur, whom I had the opportunity to work with during a short period of time.

I would like to thank professor Demkowicz for his inspiring lectures and professor Heimbach for accepting to be in my committee.

I started my PhD with the modeling of oil spill in the Gulf of Mexico, but then decided to change to the more recent topic of HDG methods. All the codes written and results obtained were done completely independently without the help of anyone in the group or from outside.

# Space-Time Hybridized Discontinuous Galerkin Methods for Shallow Water Equations

Publication No. _____

Hamidreza Arabshahi, Ph.D.
The University of Texas at Austin, 2016

Supervisor: Clinton Dawson

The non-linear shallow water equations model the dynamics of a shallow layer of an incompressible fluid; they are obtained by asymptotic analysis and depth-averaging of the Navier-Stokes equations. They are utilized in a wide range of applications, from simulation of geophysical phenomena such as river/oceanic flows and avalanches to the study of hurricane simulation, storm surge modeling, and oil spills. As a hyperbolic system of equations, shocks may develop in finite time and therefore an appropriate numerical discretization of these equations needs to be developed. The purpose of this dissertation is to develop and implement a state of the art numerical method to accurately model these equations. Therefore, a well-balanced space-time hybridized discontinuous Galerkin method was developed for our purpose. The method was implemented and tested for several benchmark problems and very promising results were obtained. An *a priori* error estimate for the developed method was also obtained with an optimal rate of convergence in an appropriate norm.

The estimate obtained is an extension of the existing *a priori* error estimates in the literature, first to the case of a system of shallow water equations, second to a hybridized mixed DG method, and third to an arbitrary degree of polynomial in time.

# Table of Contents

# Chapter 1

# Introduction

## 1.1 Literature review

The non-linear shallow water equations (SWE) model the dynamics of a shallow layer of an incompressible fluid. They are obtained by asymptotic analysis and depth-averaging of the Navier-Stokes equations and are well-suited for the simulation of geophysical phenomena, such as river and oceanic flows or avalanches. This model is also extensively used in coastal engineering for the study of near shore flows involving run-up and run-down on sloping beaches and for the design of coastal structures as well as hurricane simulation, storm surge and oil spill modeling. To allow a proper simulation of such a variety of phenomena, accurate and robust numerical methods have to be used.

Finite Volume (FV) methods were one of the first successful methods applied to SWE. Their low computational cost, local conservation and their capability to capture shocks were highly desirable [24],[28]. However they usually suffer from low accuracy and one generally needs to use some reconstruction methods to offset the low order of convergence [44].

Discontinuous Galerkin (DG) methods have been developed during the past twenty years or so. These methods are essentially a combination of the

Finite Element (FE) methods, FV methods and Riemann solvers. They can handle any type of mesh, element shape and basis functions, are ideally suited for hp-adaptivity and are highly parallelizable. It is only recently that the DG approach has been applied to the SWE and there are a growing number of studies in the literature. One of the first studies was due to Schwanenberg and Harms [61]; who used a Runge-Kutta DG (RKDG) method for simulating the dam-break problem. Later Aizinger and Dawson [1] formulated the local discontinuous Galerkin method for the 2D shallow-water equations and derived stability and *a priori* error estimates for a simplified form of the equations. Ambati and Bokhove [2] studied the space-time version of DG method for the SWE. Regarding the adaptivity, Eskilsson and Sherwin [32] and Kubatko et al. [47] among others extended the method to the hp-adaptive case.

Some researches have also applied the above mentioned methods to the simulation of shallow-water equations on the sphere. Jakob-Chien et al. [43], used the spectral transform method for shallow water simulations. These class of methods demand a high computational expense at high resolution that is associated with the computational cost of the Legendre transforms. Finite-difference approaches include those of Heikes and Randall [41] and Ronchi et al. [60]. Hybrid finite-volume methods incorporate both a finite-volume treatment of conservative variables and a finite-difference treatment of momentum and include the models of Lin and Rood [49]. Finite-element type models, including spectral-element (SE) and discontinuous-Galerkin (DG) models have been presented by, Cote and Staniforth [23], Giraldo et al. [36] and Nair et al.

[52].

Hybridized DG (HDG) methods have been recently introduced in the context of DG methods [18]. They are essentially targeting two drawbacks of the DG methods, namely their high degrees of freedom (DOF) and their inability to converge optimally (sub-optimal convergence of fluxes) or even lack of convergence (in case of using zero-order (constant) elements) for some problems. These methods use discontinuous approximations for both the solution inside each element and its trace on the element boundary. They define the local solvers by using a Galerkin method to weakly enforce the equations inside each element and define a global problem by weakly imposing the transmission conditions across the elements. Based on the author's knowledge at the time of writing this thesis, the HDG method has not yet been directly applied to the modeling of the SWE.

A new class of DG methods called discontinuous Petrov Galerkin methods (DPG) has been developed by Demkowicz and Gopalakrishnan recently [25],[26],[27],[68]. These methods fall in the category of DG methods with a basic goal in mind of automatically obtaining optimal test functions for any given trial spaces; optimal in the sense of maximizing the stability constant in the energy norm of the problem at hand. However there is no guarantee that the test functions produced cause the numerical method to be locally conservative, a crucial property that is needed for any conservation laws. Some recent methods have been developed to enforce this property [30].

## 1.2 Thesis goal

Although the application of the DG methods to the SWE have been shown to be successful, there are still some issues which are relevant in all the previous work. One is the stability of the explicit DG methods which requires a small time step such that the CFL condition is satisfied. In real applications, the scale of mesh resolution can vary in a few order of magnitudes from small mesh near the shore ($\sim 200m$) to very large mesh in the ocean ($\sim 1km$) (see Figure 1.1). This would cause the largest possible time step to be governed by the smallest scale in the mesh, resulting in a very small time step (usually of order of $1 \sim 2$ seconds at most). The simulation time is also in order of days. This would require a fair amount of computational time and resources. Therefore implicit methods might be an option to bypass the stability limitations. However the the large number of degree of freedom need to be solved (specially for higher order methods) make the method not attractable. Therefore if we can reduce the total number of DOF need to be solved, we can save a huge amount in terms of computational resource. The goal of this thesis to develop a stable and more accurate scheme for the shallow water equations via using the following two key features:

- Hybridization: This would reduce (for higher order methods) the number of DOF need to be solved in implicit method. It also increases the accuracy of the method by improving the order of convergence and we have the option of post-processing which can even make the results more accurate.

- Space-Time methods: As implicit methods in nature, they would

potentially allow for much bigger time-steps compared to explicit methods, can deal with moving meshes and naturally satisfy Geometric Conservation Law(GCL) which is crucial for the accuracy of the solution on moving meshes.

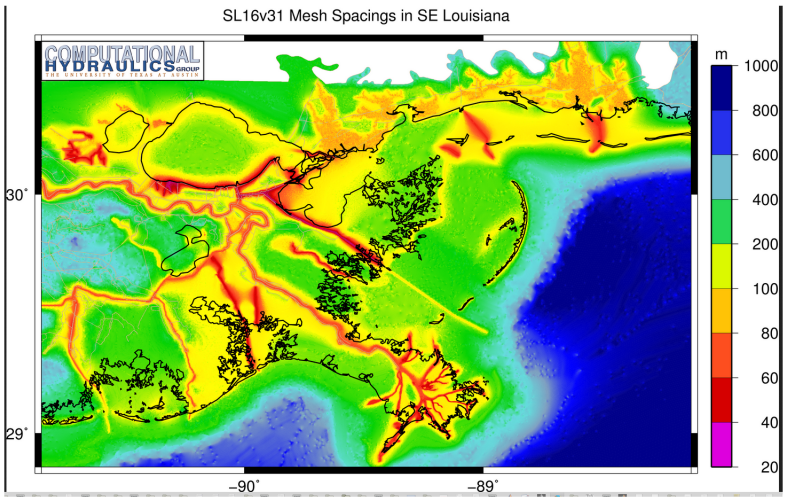


Figure 1.1: Size of elements in a typical mesh

## 1.3 Scope

The dissertation will proceed in the next six chapters. In chapter two, the SWE are derived, the basic assumptions are stated and numerical treatment of the boundary conditions will be described. In chapter three, we will introduce the overall formulation of the space-time DG (STDG) and space-time hybridized (STHDG) methods for a general system of advection-diffusion equations. In chapter four, the weak formulation of STHDG method for the SWE is derived, a well-balanced scheme will be introduced and some details

of the implementation will be stated. In chapter five the numerical results will be covered. In chapter six an *a priori* error estimate for STHDG will be obtained and at last the future work will be covered in chapter seven.

# Chapter 2

# Shallow Water Equations

The shallow water equations are utilized in different contexts in both oceanographic and atmospheric fluid flow such as modeling flow in rivers and coastal areas, tsunami prediction, climate forecast, storm surge modeling, etc. They are a two-dimensional system of hyperbolic conservation laws. In this chapter the shallow water equations are derived and the main assumptions stated.

## 2.1 Derivation

We start with the conservation laws of mass and momentum for a compressible medium, written in differential form

$$\rho_t + \nabla \cdot (\rho \vec{V}) = 0 \tag{2.1}$$

$$\frac{\partial}{\partial t}(\rho \vec{V}) + \nabla \cdot [\rho \vec{V} \otimes \vec{V} + pI - \Pi] = \rho \vec{g} \tag{2.2}$$

where $\rho$ is mass, $\vec{V} = (u, v, w)$ is the velocity vector, $p$ is pressure, $\vec{g} = (g_1, g_2, g_3)$ is a body force vector, $I$ is the unit tensor and $\Pi$ is the viscous stress tensor.

Assuming the density of the fluid is constant, $\vec{g} = (0, 0, -g)$ and ignoring the

viscosity term for now, we can expand the mass and momentum equations into the following:

$$
\begin{aligned}
&u_x + v_y + w_z = 0, \\
&u_t + uu_x + vu_y + wu_z = -\frac{1}{\rho}(p_x), \\
&v_t + uv_x + vv_y + wv_z = -\frac{1}{\rho}(p_y), \\
&w_t + uw_x + vw_y + ww_z = -\frac{1}{\rho}(p_z) - g.
\end{aligned}
\tag{2.3}
$$

We now need to introduce the boundary conditions. The bottom topography is defined by a surface as

$$z = b(x, y), \tag{2.4}$$

and the top (free surface) boundary as

$$z = s(x, y, t) = b(x, y) + h(x, y, t), \tag{2.5}$$

where $h$ is the depth of the water between the free surface and bottom topography. Note that the position of the free surface is not known *a priori*.
We now make some simplifications. First it is assumed that the vertical component of acceleration is negligible, i.e.

$$\frac{dw}{dt} = w_t + w_t + uw_x + vw_y + ww_z = 0.$$

Inserting this condition into that last equation of (2.3) and integrating in $z$, we obtain

$$p = \rho g(s - z), \tag{2.6}$$

which is the hydrostatic pressure assumption. Differentiation of (2.6) with respect to $x$ and $y$ gives

$$p_x = \rho g s_x, \quad p_y = \rho g s_y.$$

Note that $p_x, p_y$ are independent of $z$ and so are the left hand sides of the second and third equations in (2.3), i.e. the accelerations $\frac{du}{dt}$ and $\frac{dv}{dt}$. Hence the $u, v$ components of velocity would be independent of $z$. By virtue of the above conditions we can simplify the equations (2.3) as

$$\begin{aligned} &u_x + v_y + w_z = 0, \\ &u_t + uu_x + vu_y = -\frac{1}{\rho}(p_x), \\ &v_t + uv_x + vv_y = -\frac{1}{\rho}(p_y), \\ &p = \rho g(s - z). \end{aligned} \tag{2.7}$$

We now integrate the continuity equation in $z$ between the bottom ($z = b(x, y)$) and free surface ($z = s(x, y, t)$). That is

$$\int_b^s (u_x + v_y + w_z)dz = 0,$$

which leads to

$$w|_{z=s} - w|_{z=b} + \int_s^b u_x dz + \int_b^s v_y dz = 0. \tag{2.8}$$

We now evaluate the first two boundary conditions. Differentiating equation (2.5) in time, we have

$$(w - us_x + vs_y + s_t)|_{z=s} = 0,$$

and the same for (2.4), we obtain

$$(w - ub_x + vb_y)|_{z=b} = 0.$$

Substituting in (2.8), we obtain

$$(us_x + vs_y + s_t)|_{z=s} - (ub_x + vb_y)|_{z=b} + \int_s^b u_x dz + \int_b^s v_y dz = 0. \tag{2.9}$$

Now the last two integral terms can be simplified using Leibniz's rule for differentiation under the integral sign. Therefore

$$\int_s^b u_x dz = \frac{\partial}{\partial x} \int_s^b u dz - u|_{z=s} s_x + u|_{z=b} b_x \tag{2.10}$$

and

$$\int_s^b v_y dz = \frac{\partial}{\partial y} \int_s^b v dz - v|_{z=s} s_y + v|_{z=b} b_y. \tag{2.11}$$

Inserting (2.10) and (2.11) into (2.9), we obtain

$$s_t + \frac{\partial}{\partial x} \int_b^s u dz + \frac{\partial}{\partial y} \int_b^s v dz = 0. \tag{2.12}$$

Recalling that $u$ and $v$ are independent of $z$, $s = b + h$ and $b_t = 0$, equation (2.12) can be simplified as

$$h_t + (hu)_x + (hv)_y = 0, \tag{2.13}$$

which is the law of conservation of mass written in differential form.
To obtain the momentum equations, we need the following assumption

$$h\frac{\partial h}{\partial x} = \frac{1}{2}\frac{\partial h^2}{\partial x}. \tag{2.14}$$

Although this looks straightforward, we are actually replacing a non-conservative product with a conservative one. The left hand side is a multiplication of a distribution (a discontinuous function in general) with a measure (a delta function), which is mathematically not defined. However the right hand side is a well defined term. The non-conservative products have been a major source of error in numerical simulation of shallow water equations and several approximating theories have been developed to deal with these terms such as Colombeau algebra [22], or the more famous work of Dal Maso, LeFloch, and Murat [51] based on the assumption that the path connecting the discontinuities are known *a priori* and is a straight line in this case. The problem is that the path is unknown, although some results have been obtained for the case of a simple shock [67].

Using the assumption (2.14), pre-multiplying equation (2.13) by $u$ and the second equation in (2.7) by $h$ and adding the results together, we obtain

$$(hu)_t + (hu^2 + \frac{1}{2}gh^2)_x + (huv)_y = -ghb_x, \tag{2.15}$$

and similarly for the $y$ momentum, we get

$$(hv)_t + (huv)_x + (hv^2 + \frac{1}{2}gh^2)_y = -ghb_y. \tag{2.16}$$

Note that the right hand side terms are still in non-conservative form. The three equations (2.13), (2.15) and (2.16) can be written compactly as

$$\frac{\partial}{\partial t}\vec{U} + \frac{\partial}{\partial x}\vec{F}(\vec{U}) + \frac{\partial}{\partial y}\vec{G}(\vec{U}) = \vec{S}(\vec{U}) \tag{2.17}$$

where

$$\vec{U} = \begin{pmatrix} h \\ hu \\ hv \end{pmatrix}, \qquad \vec{S} = \begin{pmatrix} 0 \\ -ghb_x \\ -ghb_y \end{pmatrix}$$
$$\vec{F}(\vec{U}) = \begin{pmatrix} hu \\ hu^2 + \frac{1}{2}gh^2 \\ huv \end{pmatrix}, \quad \vec{G}(\vec{U}) = \begin{pmatrix} hv \\ huv \\ hv^2 + \frac{1}{2}gh^2 \end{pmatrix}. \tag{2.18}$$

$\vec{U}$ is generally called the vector of conserved variable, $\vec{F}(\vec{U})$ and $\vec{G}(\vec{U})$ are fluxes in the $x$ and $y$ direction respectively, $\vec{S}(\vec{U})$ is the source term which in general contains other terms such as the Coriolis forces, wind forces and bottom friction.

## 2.2 Characteristic form

The characteristic form of (2.17) can be obtained by introducing the celerity $c = \sqrt{gh}$ of pressure waves in still water. Denoting by $A$ and $B$ the Jacobian matrices of $F$ and $G$ respectively, equations (2.17) can be written as

$$\frac{\partial}{\partial t}\vec{U} + A\frac{\partial}{\partial x}\vec{U} + B\frac{\partial}{\partial y}\vec{U} = \vec{S}(\vec{U}), \tag{2.19}$$

where

$$A = \begin{pmatrix} 0 & 1 & 0 \\ c^2 - u^2 & 2u & 0 \\ -uv & v & u \end{pmatrix}, \quad B = \begin{pmatrix} 0 & 0 & 1 \\ -uv & v & u \\ c^2 - v^2 & 0 & 2v \end{pmatrix}. \tag{2.20}$$

The system of conservation laws in (2.17) is said to be hyperbolic if the matrix $M = \beta_1 A + \beta_2 B$ is diagonalizable and has three real eigenvalues for any combination of real coefficients $\beta_1, \beta_2$. Choosing $\beta_1 = n_x$ and $\beta_2 = n_y$, where $(n_x, n_y)$ represents a unit normal to a surface, the eigenvalues of the Jacobian of the normal flux can be obtained as

$$\begin{aligned} \lambda_1 &= un_x + vn_y + c \\ \lambda_2 &= un_x + vn_y \\ \lambda_3 &= un_x + vn_y - c, \end{aligned} \tag{2.21}$$

and the corresponding eigenvectors are

$$e_1 = \begin{pmatrix} 1 \\ u + cn_x \\ v + cn_y \end{pmatrix}, \quad e_2 = \begin{pmatrix} 0 \\ -cn_y \\ cn_x \end{pmatrix} \quad e_3 = \begin{pmatrix} 1 \\ u - cn_x \\ v - cn_y \end{pmatrix}.$$

Based on these eigenvectors, a matrix, $P$ can be constructed such that they make the Jacobian of the normal flux diagonalizable, i.e.

$$An_x + Bn_y = P\Lambda P^{-1},$$

where

$$P = \begin{pmatrix} 1 & 0 & 1 \\ u + cn_x & -cn_y & u - cn_x \\ v + cn_y & cn_x & v - cn_y \end{pmatrix}, \quad P^{-1} = \frac{1}{2c} \begin{pmatrix} un_x + vn_y + c & nx & ny \\ 2(un_y - vn_x) & -2n_y & 2n_x \\ un_x + vn_y + c & -n_x & -n_y \end{pmatrix},$$

and

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix}.$$

## 2.3 Boundary conditions

For hyperbolic equations, the number of boundary conditions prescribed at a boundary depends on the number of characteristics (waves) entering the computational domain. In the context of shallow water equations, this method of characteristics has been widely used as an approximation to impose the boundary conditions. Essentially if a characteristic enters the domain, a boundary condition needs to be specified and if it leaves the domain, no boundary condition is needed. Hence the basic idea is to check the sign of the eigenvalues obtained in (2.21). For an inflow boundary, as $un_x + vn_y < 0$, ($h > 0$) two of the eigenvalues are negative, i.e. $\lambda_2 < 0$ and $\lambda_3 < 0$ and therefore their corresponding waves are entering the domain. Now if $un_x + vn_y \leq -c$, the third wave is also entering and three boundary conditions need to be imposed, otherwise two boundary conditions are needed. In the first case the flow is called supercritical and in the second case subcritical. This scenario

can be repeated for outflow boundary conditions as well. In summary

| | |
|---|---|
| $un_x + vn_y \leq -c$ | supercritical inflow |
| $-c \leq un_x + vn_y \leq 0$ | subcritical inflow |
| $0 \leq un_x + vn_y \leq c$ | subcritical outflow |
| $un_x + vn_y \geq c$ | supercritical outflow. |

# Chapter 3

# An Overview of DG and HDG Methods

The classical finite element methods have been proved to successfully model different types of equations, but when it comes to conservation laws and generally hyperbolic equations, they mostly suffer from two major drawbacks; one is stability and the second one is their inability to satisfy local conservation. To resolve some of these issues, discontinuous Galerkin (DG) methods were therefore proposed as an extension of total variation diminishing (TVD) and total variation bounded (TVB) finite difference methods for hyperbolic conservation laws. These methods are a class of non conforming finite element methods whereby the underlying constraints (i.e. conformity) in the finite element space are removed and imposed weakly in the variational formulation. The removal of these constraints introduces an additional flexibility in terms of modeling, which can be taken advantage of in numerical simulation. This flexibility comes with the price of additional degrees of freedom and thus more computational time compared to classical finite elements. A remedy for this has been recently introduced in the context of DG methods (even though the idea goes back to classical finite element analysis) in which additional degrees of freedom are introduced at element interfaces. These new degrees of freedom (DOF) act two-fold. On one hand they are used to locally condensate the DOF

inside an element to those on the boundary and thus reduce the total number of equations needed to be solved, and second they act as Lagrange multipliers to enforce the constraint which was originally removed from the underlying finite element space. This reduction in number of DOF would make implicit methods such as space-time finite element methods more attractive. As a result we are using DG methods both in space and time and hybridization in space, resulting in space-time hybridized discontinuous Galerkin method (STHDG).

In this chapter an overview of space-time DG (STDG) and STHDG methods are given for a general advection diffusion problem. The specific implementation for shallow water equations is covered in the next chapter.

## 3.1 Space-time DG methods

We introduce some notation which is used in the following to define the method. Let $\Omega \subset \mathbb{R}^d, d = 2, 3$ be a bounded domain, $T > 0$ and $Q_T = \Omega \times (0, T)$. The time discretization is based on a partition of $0 = t_0 < t_1 < ... < t_M = T$ of the time interval $I = [0, T]$ into subintervals $I_m = (t_{m-1}, t_m)$ of length $\tau_m = t_m - t_{m-1}$. Each space-time slab, $S_m = \Omega \times I_m$ is divided into space-time prisms $K_i \times I_m$, where $\mathfrak{T}_{h,m} = \{K_i\}, i \in I$ is a triangulation of $\Omega$ with mesh discretization parameter $h_m$ and $I$ is an index set. The space mesh may change from one time interval to another (e.g. due to adaptivity). Thus in general there are two space-meshes associated to each time level $t_m$, namely one from below and one from above. In case that these two meshes are

not aligned, hanging nodes are inevitable. An interior face "$e$" is any planar set of positive $(d-1)$-dimensional measure of the form $e = \partial K^+ \cap \partial K^-$ for some two elements $K^+, K^- \in \mathcal{T}_{h,m}$. Similarly, we say that "$e$" is a boundary face if $e = \partial K^+ \cap \partial\Omega$ and the $(d-1)$-dimensional Lebesgue measure of "$e$" is not zero. We define $\Gamma_{h,m} = \{e\}_i, i \in J$ as the set of all faces of the mesh where $J$ is an index set. By $\Gamma^{int}_{h,m} = \{e\}_i, i \in J^{int}$ and $\Gamma^{b}_{h,m} = \{e\}_i, i \in J^{b}$ we mean the set of all interior and boundary faces, respectively. Therefore $J = J^{int} \cup J^b$. The set of boundary faces is also defined to be the union of Dirichlet ($\Gamma^{b,D}_{h,m} = \{e\}_i, i \in J^b_D$) and Neumann ($\Gamma^{b,N}_{h,m} = \{e\}_i, i \in J^b_N$) parts of the boundary. We define $\Xi_T = \bigcup_m \Gamma_{h,m} \times I_m$ and we also have $Q_T = \bigcup_m \mathcal{T}_{h,m} \times I_m$. As the solution is discontinuous between time slabs we also define

$$\phi^{\pm}_m = \lim_{\epsilon>0, \epsilon\to 0} \phi(t_m \pm \epsilon), \quad [\![\phi]\!]_m = \phi^+_m - \phi^-_m.$$

for an arbitrary function $\phi$.

### 3.1.1 Finite element spaces

The usual finite element space for DG methods is the broken Sobolev spaces defined for any integer $k \geq 1$ as

$$H^k(\mathcal{T}_{h,m}) = \{v \in L^2(\Omega) : v|_K \in H^k(K)\}.$$

As we are working with space-time method, we essentially need to use the broken Bochner spaces (vector-valued function). We first recall the definition of these spaces as following [33]:

$$L^p(I;X) = \{f : \|f\|_{L^p(I;X)} := \|\|f\|_X\|_{L^p(I;\mathbb{R})} = \Big(\int_I \|f(t)\|^p_X \, dt\Big)^{\frac{1}{p}} < \infty\},$$

$$H^k(I;X) = \{f : \frac{\partial^t f}{\partial x^t} \in L^p(I;X), t = 0, ..., k, \|f\|_{H^k(I;X)} < \infty\}$$

where $X$ is a Banach space and $I$ an open interval and

$$\|f\|_{H^k(I;X)} = \Big( \int_I \sum_{t=0}^{k} \Big\| \frac{\partial^t f(t)}{\partial x^t} \Big\|_X^p dt \Big)^{\frac{1}{p}}.$$

The derivatives are in the sense of distributions. The broken Bochner space can be defined as

$$H^k(I_m; H^p(\mathcal{T}_{h,m})) = \{v : \Omega \times I_m \to \mathbb{R} : v|_{K \times I_m} \in H^k(I_m; H^p(K))\},$$

for $p, k \geq 1$.

For the finite-dimensional setting, we define the following two spaces

$$S^p_{h,m} = \{\phi \in L^2(\mathcal{T}_{h,m}) : \phi|_K \in \mathcal{P}^P(K)\}$$

$$S^{p,q}_{h,\tau} = \{\phi \in L^2(Q_T) : \phi|_{I_m \times K} \in \mathcal{P}^q(I_m; \mathcal{P}^p(K))\},$$

where $\mathcal{P}^q$ denotes the space of polynomials of order less than or equal to $q$. $S^p_{h,m}$ and $S^{p,q}_{h,\tau}$ are finite dimensional versions of broken Sobolev and Bochner spaces, respectively.

### 3.1.2 Space-time DG discretization for a general system of advection diffusion problem

Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with a piecewise smooth Lipschitz continuous boundary $\partial\Omega$ and let $T > 0$. In a space-time cylinder $Q_T = \Omega \times (0, T)$, the problem is to find the vector-valued function $\vec{w} = (w_1, ..., w_l)^T : Q_T \to \mathbb{R}^l$ such that

$$\frac{\partial \vec{w}}{\partial t} + \sum_{s=1}^{d} \frac{\partial \vec{f}_s(\vec{w})}{\partial x_s} - \nabla \cdot (\mathbb{K}(\vec{w}) \nabla \vec{w}) = \vec{F}(\vec{w})$$

$$\vec{w}(\vec{x},0) = \vec{w}^0(\vec{x}), \quad \vec{x} \in \Omega$$

$$\vec{w}(\vec{x},t) = \vec{g}_D, \quad (\vec{x},t) \in \partial\Omega_D \times (0,t)$$

$$(\mathbb{K}(\vec{w})\nabla\vec{w}) \cdot \vec{n} = \vec{g}_N, \quad (\vec{x},t) \in \partial\Omega_N \times (0,t) \tag{3.1}$$

where $l$ is the number of equations in the system, $d$ the dimension, $\vec{g}_D, \vec{g}_N$ are the boundary conditions defined on the Dirichlet ($\partial\Omega_D$) and Neumann ($\partial\Omega_N$) part of the boundary, respectively. $\vec{w}$ is the vector of state variables, $\vec{f}_s(\vec{w})$ is the convective flux of the vector $\vec{w}$ in direction $s$, $\vec{n}$ is the unit normal to the boundary, $\mathbb{K}$ is the diffusion matrix and $\vec{F}$ corresponds to source terms. The exact function spaces needed to define a well-posed problem generally depends of the type of the convective flux used in the equation. However based on the time derivative term and the diffusion term we need to assume that the solution is at least $\vec{w} \in (C^0(H^2(\Omega)))^l \cap (C^1(L^2(\Omega)))^l$ which results in the initial condition to be $\vec{w}(\vec{x},0) \in (H^1(\Omega))^l$.

#### 3.1.2.1 Weak form

To discretize, we multiply (3.1) by functions $\vec{\phi} \in (H^k(I_m; H^p(\mathcal{T}_{h,m})))^l$, integrate over an element $K_i \times I_m$ and then sum over all space-time elements to obtain

$$\begin{aligned}
&\int_{I_m} \sum_{i\in I} \int_{K_i} (\frac{\partial \vec{w}}{\partial t} \cdot \vec{\phi}) dxdt + \int_{I_m} (\sum_{i\in I} \int_{K_i} \sum_{s=1}^{d} \frac{\partial \vec{f}_s(\vec{w})}{\partial x_s} \cdot \vec{\phi}) dxdt \\
&- \int_{I_m} (\sum_{i\in I} \int_{K_i} \nabla \cdot (\mathbb{K}(\vec{w})\nabla\vec{w}) \cdot \vec{\phi}) dxdt \\
&= \int_{I_m} (\sum_{i\in I} \int_{K_i} \vec{F}(\vec{w}) \cdot \vec{\phi}) dxdt \quad \forall \vec{\phi} \in (H^k(I_m; H^p(\mathcal{T}_{h,m})))^l.
\end{aligned}$$

Now each term on the left hand side can be discretized as follows:
Utilizing the Green theorem, the time derivative term can be written as

$$\int_{I_m} \sum_{i\in I} \int_{K_i} (\frac{\partial \vec{w}}{\partial t} \cdot \vec{\phi}) dxdt = \sum_{i\in I} \int_{K_i} (\hat{\vec{w}} \cdot \vec{\phi} n_t) dx - \int_{I_m} \sum_{i\in I} \int_{K_i} \vec{w} \cdot (\frac{\partial \vec{\phi}}{\partial t}) dxdt$$

where $\hat{w}$ is the trace values of solution at the time level $t^+_{m-1}$ where $n_t = -1$ and $t^-_m$ where $n_t = 1$. For the numerical flux, an upwinding in time is chosen. This is the most straight forward choice of flux as essentially all the phenomena advect in time from past to the future. By performing another integration by parts we obtain the following terms for time integration

$$\sum_{i\in I} \int_{K_i} (\vec{w}(., t^+_{m-1}) - \vec{w}^{\uparrow}) \cdot \vec{\phi} dx + \int_{I_m} \sum_{i\in I} \int_{K_i} (\frac{\partial \vec{w}}{\partial t} \cdot \vec{\phi}) dxdt \tag{3.2}$$

where $\vec{w}^{\uparrow} := \vec{w}(., t^-_{m-1})$ is the value upwinded from the previous time step. Before moving to the next term a caveat need to be mentioned. If the triangulation is different for $t^-_{m-1}$ and $t^+_{m-1}$, then a pure upwind is not feasible as it introduces discontinuity inside the space-time element of the next time slab and therefore the solution $\vec{w}$ will not be in further in $(S^{p,q}_{h,\tau})^l$. In this case the upwinded solution must be first $L^2$ projected onto the space $S^p_{h,m}$ at time $t^+_{m-1}$ and then the procedure can be continued.

Now for the second term the advective flux can be integrated by parts and we obtain

$$\int_{I_m} (\sum_{i\in I} \int_{K_i} \sum_{s=1}^{d} \frac{\partial \vec{f}_s(\vec{w})}{\partial x_s} \cdot \vec{\phi}) dxdt =$$
$$\int_{I_m} (\sum_{i\in I} \int_{\partial K_i} \sum_{s=1}^{d} \hat{\vec{f}}_s(\vec{w}) \cdot \vec{\phi} n_s) dsdt - \int_{I_m} (\sum_{i\in I} \int_{K_i} \sum_{s=1}^{d} \vec{f}_s \cdot \frac{\partial \vec{\phi}}{\partial x_s}) dxdt.$$

The convective flux in the boundary term can be substituted by the numerical flux $\vec{H}(\vec{w}^+, \vec{w}^-, n_s)$, with the properties of consistency and conservativity. $\vec{w}^+, \vec{w}^-$ are the left and right states of an element edge. Inserting the numerical flux, we have

$$\int_{I_m}(\sum_{i\in I}\int_{K_i}\sum_{s=1}^{d}\frac{\partial \vec{f}_s(\vec{w})}{\partial x_s}\cdot\vec{\phi})dxdt = \int_{I_m}(\sum_{i\in I}\int_{\partial K_i\setminus\partial\Omega}\sum_{s=1}^{d}\vec{H}(\vec{w}^+,\vec{w}^-,n_s)\cdot\vec{\phi}n_s)dsdt$$
$$+\int_{I_m}(\sum_{i\in I}\int_{\partial K_i\cap\partial\Omega_D}\sum_{s=1}^{d}\vec{H}(\vec{g}_D,\vec{w}^-,n_s)\cdot\vec{\phi}n_s)dsdt$$
$$+\int_{I_m}(\sum_{i\in I}\int_{\partial K_i\cap\partial\Omega_N}\sum_{s=1}^{d}\vec{H}(\vec{w}^-,\vec{w}^-,n_s)\cdot\vec{\phi}n_s)dsdt$$
$$-\int_{I_m}(\sum_{i\in I}\int_{K_i}\sum_{s=1}^{d}\vec{f}_s\cdot\frac{\partial\vec{\phi}}{\partial x_s})dxdt. \tag{3.3}$$

where for the Neumann part of the boundary conditions, for evaluating the numerical flux, we have used an extrapolation of the $\vec{w}$ from inside.

For the diffusion term, after using the Green theorem we obtain

$$\int_{I_m}(\sum_{i\in I}\int_{K_i}\nabla\cdot(\mathbb{K}(\vec{w})\nabla\vec{w})\cdot\vec{\phi})dxdt = -\int_{I_m}(\sum_{i\in I}\int_{K_i}(\mathbb{K}(\vec{w})\nabla\vec{w}):\nabla\vec{\phi})dxdt$$
$$+\int_{I_m}(\sum_{i\in I}\int_{\partial K_i}(\mathbb{K}(\vec{w})\nabla\vec{w}):(\vec{\phi}\otimes\vec{n}))dsdt. \tag{3.4}$$

In order to proceed, we need to define the average and jump operator for scalar quantity $V$, vector $\vec{V}$ and tensor quantity $\mathbb{V}$ as follows:

$$\langle V\rangle = \frac{1}{2}(V^+ + V^-),\ \ \langle\vec{V}\rangle = \frac{1}{2}(\vec{V}^+ + \vec{V}^-),\ \ \langle\mathbb{V}\rangle = \frac{1}{2}(\mathbb{V}^+ + \mathbb{V}^-),$$

$$[\![V]\!] = V^+\vec{n}^+ + V^-\vec{n}^-,\ \ [\![\vec{V}]\!] = \vec{V}^+\otimes\vec{n}^+ + \vec{V}^-\otimes\vec{n}^-,\ \ [\![\mathbb{V}]\!] = \mathbb{V}^+\vec{n}^+ + \mathbb{V}^-\vec{n}^-.$$

We have the following simple identity for the jump of a product term

$$[\![ab]\!] = [\![a]\!]\langle b\rangle + [\![b]\!]\langle a\rangle.$$

Now using this identity, the boundary integral in (3.4) can be written as

$$\begin{aligned}
&\int_{I_m} (\sum_{i\in I} \int_{\partial K_i} (\mathbb{K}(\vec{w})\nabla\vec{w}) : (\vec{\phi}\otimes\vec{n}))dsdt = \\
&\int_{I_m} (\sum_{i\in I} \int_{\partial K_i\setminus\partial\Omega} (\mathbb{K}(\vec{w})\nabla\vec{w}) : (\vec{\phi}\otimes\vec{n}))dsdt \\
&+ \int_{I_m} (\sum_{i\in I} \int_{\partial K_i\cap\partial\Omega} (\mathbb{K}(\vec{w})\nabla\vec{w}) : (\vec{\phi}\otimes\vec{n}))dsdt \\
&= \int_{I_m} (\sum_{e\in\Gamma_{h,m}^{int}} \int_e \langle \mathbb{K}(\vec{w})\nabla\vec{w}\rangle : [\![\vec{\phi}]\!])dsdt + \int_{I_m} (\sum_{e\in\Gamma_{h,m}^{int}} \int_e [\![\mathbb{K}(\vec{w})\nabla\vec{w}]\!]\cdot\langle\vec{\phi}\rangle)dsdt \\
&+ \int_{I_m} (\sum_{e\in\Gamma_{h,m}^{b,D}} \int_e (\mathbb{K}(\vec{w})\nabla\vec{w})\cdot\vec{n}\cdot\vec{\phi})dsdt + \int_{I_m} (\sum_{e\in\Gamma_{h,m}^{b,N}} \int_e \vec{g}_N\cdot\vec{\phi})dsdt
\end{aligned}$$

As we have assumed that $\vec{w}\in(C^0(H^2(\Omega)))^l\cap(C^1(L^2(\Omega)))^l$ then if the matrix $\mathbb{K}$ is regular enough, the product term $\mathbb{K}(\vec{w})\nabla\vec{w}$ is in $(H(div;\Omega))^l$ space which means that its normal component must be continuous across elements, i.e. $[\![\mathbb{K}(\vec{w})\nabla\vec{w}]\!] = 0$. We will now add two additional terms (Nitsche's terms [55]) to the weak form

$$\pm\int_{I_m} (\sum_{e\in\Gamma_{h,m}} \int_e \langle \mathbb{K}(\vec{w})\nabla\vec{\phi}\rangle : [\![\vec{w}]\!])dsdt + \int_{I_m} (\sum_{e\in\Gamma_{h,m}} \int_e \beta[\![\vec{w}]\!] : [\![\vec{\phi}]\!])dsdt.$$

The last term is the penalty term to weakly impose the continuity across the elements. The effect of the first term can be best understood when the diffusion matrix $\mathbb{K}$ does not depend on the solution. Then if it is used with minus sign, it would symmetrize the weak form. This is desirable as the continuous setting

(the diffusion equation) is symmetric too and makes the weak form adjoint consistent . However, in this case the penalty parameter $\beta$ should be taken large enough so that we can guarantee the stability. If used with plus sign, the weak form corresponding to diffusion would not be symmetric but it would be unconditionally stable. This can be easily seen by choosing $\vec{w} = \vec{\phi}$, in the weak form, thus two of the terms cancel out and we end up with positive terms only which can be taken as a norm in the finite dimensional setting.

In summary we will obtain the following weak form for the STDG method:

Find $\vec{w} \in (H^k(I_m; H^p(\mathcal{T}_{h,m})))^l$ such that

$$\begin{aligned}
&\int_{I_m} \sum_{i\in I} \int_{K_i} (\frac{\partial \vec{w}}{\partial t} \cdot \vec{\phi}) dxdt + \sum_{i\in I} \int_{K_i} (\vec{w}(.,t_{n-1}^+) - \vec{w}^{\uparrow}) \cdot \vec{\phi} dx \\
&+ \int_{I_m} (\sum_{i\in I} \int_{\partial K_i \setminus \partial\Omega} \sum_{s=1}^{d} \vec{H}(\vec{w}^+, \vec{w}^-, n_s) \cdot \vec{\phi} n_s) dsdt \\
&+ \int_{I_m} (\sum_{i\in I} \int_{\partial K_i \cap \partial\Omega_D} \sum_{s=1}^{d} \vec{H}(\vec{g}_D, \vec{w}^-, n_s) \cdot \vec{\phi} n_s) dsdt \\
&+ \int_{I_m} (\sum_{i\in I} \int_{\partial K_i \cap \partial\Omega_N} \sum_{s=1}^{d} \vec{H}(\vec{w}^-, \vec{w}^-, n_s) \cdot \vec{\phi} n_s) dsdt \\
&- \int_{I_m} (\sum_{i\in I} \int_{K_i} \sum_{s=1}^{d} \vec{f}_s \cdot \frac{\partial \vec{\phi}}{\partial x_s}) dxdt \\
&+ \int_{I_m} (\sum_{i\in I} \int_{K_i} (\mathbb{K}(\vec{w}) \nabla \vec{w}) : \nabla \vec{\phi}) dxdt \\
&- \int_{I_m} (\sum_{e\in \Gamma_{h,m}^{int}} \int_e \langle \mathbb{K}(\vec{w}) \nabla \vec{w} \rangle : [\![\vec{\phi}]\!]) dsdt \\
&- \int_{I_m} (\sum_{e\in \Gamma_{h,m}^{b,D}} \int_e (\mathbb{K}(\vec{w}) \nabla \vec{w}) \cdot \vec{n} \cdot \vec{\phi}) dsdt - \int_{I_m} (\sum_{e\in \Gamma_{h,m}^{b,N}} \int_e \vec{g}_N \cdot \vec{\phi}) dsdt \\
&\pm \int_{I_m} (\sum_{e\in \Gamma_{h,m}^{int}} \int_e \langle \mathbb{K}(\vec{w}) \nabla \vec{\phi} \rangle : [\![\vec{w}]\!]) dsdt + \int_{I_m} (\sum_{e\in \Gamma_{h,m}^{int}} \int_e \beta [\![\vec{w}]\!] : [\![\vec{\phi}]\!]) dsdt \\
&\pm \int_{I_m} (\sum_{e\in \Gamma_{h,m}^{b,D}} \int_e \mathbb{K}(\vec{w}) \nabla \vec{\phi} \cdot \vec{n} \cdot \vec{g}_D) dsdt + \int_{I_m} (\sum_{e\in \Gamma_{h,m}^{b,D}} \int_e \beta \vec{g}_D \cdot \vec{\phi}) dsdt \\
&\pm \int_{I_m} (\sum_{e\in \Gamma_{h,m}^{b,N}} \int_e \mathbb{K}(\vec{w}) \nabla \vec{\phi} \cdot \vec{n} \cdot \vec{w}) dsdt + \int_{I_m} (\sum_{e\in \Gamma_{h,m}^{b,N}} \int_e \beta \vec{w} \cdot \vec{\phi}) dsdt \\
&= \int_{I_m} (\sum_{i\in I} \int_{K_i} \vec{F}(\vec{w}) \cdot \vec{\phi}) dxdt \quad \forall \vec{\phi} \in (H^k(I_m; H^p(\mathcal{T}_{h,m})))^l. \qquad (3.5)
\end{aligned}$$

The weak form corresponding to the finite dimensional setting of (3.5)

will be defined by substituting $\vec{w}$ and $\vec{\phi}$ with $\vec{w}_h$ and $\vec{\phi}_h$ respectively where $\vec{w}_h, \vec{\phi}_h \in (S^P_{h,m})^l \subset (H^k(I_m; H^p(\mathfrak{T}_{h,m})))^l$ and will not be repeated here.

We will now expand the method to hybridized DG (HDG). First a brief overview of the HDG method is given and then the procedure is explained for a general advection diffusion problem.

## 3.2 HDG methods

A finite element method is a hybrid method if it involves the simultaneous approximation of a vector field defined on the union of the elements of the discretization and another one defined on the union of the boundaries of the elements. The pioneering works are due to Pian [56] and Fraejis de Veubeke [65] in the context of linear elasticity problems. The mathematical analysis was developed by Raviart and Thomas [58]. These results were further advanced by a post-processing technique developed by Arnold and Brezzi[3]. In the context of DG methods, this technique was revived first in the framework of diffusion problem by Cockburn, Gopalakrishnan and Lazarov in [18]. The method was compared to the well established methods of Raviart-Thomas (RT) [57] and Brezzi-Douglas-Marini (BDM) [9] and it was shown that these two methods can be obtained as particular cases of the new HDG method developed. This essentially means that the HDG method developed, can achieve optimal order of convergence for all the unknowns along with an efficient implementation. The method was then expanded to the case of steady-state diffusion [16], time-dependent diffusion [13], the wave equation [20], convection-diffusion equation

[17],[54],[53], linear and nonlinear elasticity [62], Stokes flow [19], incompressible Navier-Stokes [12] among others. In this section we extend the weak formulation obtained in Section 2.1 to the hybrid methods. We are essentially using a hybrid-mixed method for our formulation. A finite element is a mixed method if it involves the simultaneous approximation of two or more vector fields defined on the physical domain.

### 3.2.1 Space-time HDG discretization for a general system of advection diffusion problem

Consider again the set of equations (3.1). We first need to extend the defined function spaces to the trace functions corresponding to elements' boundaries. The functions in trace space need not to be continuous across the time slab, so a weaker regularity in time can be used. The broken Bochner space for the trace elements are defined as:

$$H^k(I_m; H^p(\Gamma_{h,m})) = \{v \in L^2(\Xi_T) : v|_{e\times I_m} \in H^k(I_m; H^p(e))\},$$
$$H^k(I_m; H^p(\Gamma_{h,m}))(g_D) = \{v \in H^k(I_m; H^p(\Gamma_{h,m})) : v|_{\Gamma^{b,D}_{h,m}} = g_D\},$$

for $p, k \geq 1$. We also define the corresponding finite dimensional Bochner spaces as follows:

$$M^{p,q}_{h,\tau} = \{\phi \in L^2(\Xi_T) : \phi|_{I_m\times e} \in \mathcal{P}^q(I_m; \mathcal{P}^p(e))\},$$
$$M^{p,q}_{h,\tau}(g_D) = \{\phi \in M^{p,q}_{h,\tau} : \phi|_{\Gamma^{b,D}_{h,m}} = Pg_D,\ \Gamma^{b,D}_{h,m} \subset \Gamma^b_{h,m}\},$$

where the operator $P$ denotes the $L^2$ projection.

#### 3.2.1.1 Weak form

Here we follow the same approach, i.e. multiplying by a test function in an appropriate space, integrating over a space-time element $K_i \times I_m$ and then summing over all elements. We are not using hybridization in the time direction. As was mentioned before hybridization in time would not provide us with any further gain as the time marching is an upwind process. However computationally it would allow us to couple the time slabs together and solving for a whole system which would then result in an extensive amount of computation. We chose not to couple the time slabs but solve for each time slab separately. So the time discretization term in (3.2) will be intact.

We now move to the flux term. As can be seen the numerical flux function $\vec{H}(\vec{w}^+, \vec{w}^-, n_s)$, couples the left and right states across an edge of an element. As we are going to introduce new unknowns (Lagrange multipliers) on element faces, the basic idea is, instead of coupling the left and right states, to couple each state with the new DOF introduced. Thus the numerical flux can be written as $\vec{H}(\vec{w}, \vec{\lambda}, n_s)$, where $\lambda$'s are the Lagrange multipliers and $\vec{w}$ corresponds to the element that we are currently computing. As an example we consider the Lax-Friedrichs flux, i.e.

$$\vec{H}(\vec{w}^+, \vec{w}^-, n_s) = \frac{1}{2}(\vec{f_s}(\vec{w}^+) + \vec{f_s}(\vec{w}^-)) + \mathbb{B}_{ad}(\vec{w}^+ - \vec{w}^-)n_s,$$

where $\mathbb{B}_{ad}$ is a local stabilization matrix. Then the hybridized numerical flux can be written as

$$\vec{H}(\vec{w}, \vec{\lambda}, n_s) = \vec{f_s}(\vec{\lambda}) + \mathbb{B}_{ad}(\vec{w} - \vec{\lambda})n_s, \tag{3.6}$$

or

$$\vec{H}(\vec{w},\vec{\lambda},n_s)=\vec{f_s}(\vec{w})+\mathbb{B}_{ad}(\vec{w}-\vec{\lambda})n_s. \tag{3.7}$$

The first form is computationally more attractable as the convective part of the flux is single-valued and thus would be canceled out in the transmission condition as will be shown later. Therefore, the weak form corresponding to (3.3) can be written as

$$\int_{I_m}(\sum_{i\in I}\int_{K_i}\sum_{s=1}^{d}\frac{\partial\vec{f_s}(\vec{w})}{\partial x_s}\cdot\vec{\phi})dxdt=\int_{I_m}(\sum_{i\in I}\int_{\partial K_i\setminus\partial\Omega_D}\sum_{s=1}^{d}\vec{H}(\vec{w},\vec{\lambda},n_s)\cdot\vec{\phi}n_s)dsdt$$
$$+\int_{I_m}(\sum_{i\in I}\int_{\partial K_i\cap\partial\Omega_D}\sum_{s=1}^{d}\vec{H}(\vec{w},\vec{g}_D,n_s)\cdot\vec{\phi}n_s)dsdt-\int_{I_m}(\sum_{i\in I}\int_{K_i}\sum_{s=1}^{d}\vec{f_s}\cdot\frac{\partial\vec{\phi}}{\partial x_s})dxdt. \tag{3.8}$$

For the diffusion part, i.e.

$$-\nabla\cdot(\mathbb{K}(\vec{w})\nabla\vec{w})=\vec{F}(\vec{w}),$$

we are going to use a mixed method as follows:

$$\mathbb{K}(\vec{w})\nabla\vec{w}+\mathbb{Z}=0,$$
$$\nabla\cdot\mathbb{Z}=\vec{F}(\vec{w}).$$

Inverting the diffusion tensor $\mathbb{K}$, multiplying by test functions and doing inte-

gration by parts we will have

$$\int_{I_m} \sum_{i \in I} \int_{K_i} (\mathbb{K}^{-1}\mathbb{Z} : \mathbb{V}) dxdt + \int_{I_m} \sum_{i \in I} \int_{\partial K_i} (\hat{\vec{w}} \otimes \vec{n} : \mathbb{V}) dsdt$$
$$- \int_{I_m} \sum_{i \in I} \int_{K_i} \vec{w} \cdot (\nabla \cdot \mathbb{V}) dxdt = 0,$$
$$- \int_{I_m} \sum_{i \in I} \int_{K_i} (\mathbb{Z} : \nabla \otimes \vec{\phi}) dxdt + \int_{I_m} \sum_{i \in I} \int_{\partial K_i} (\hat{\mathbb{Z}} \cdot \vec{n}) \cdot \vec{\phi} dsdt$$
$$= \int_{I_m} \sum_{i \in I} \int_{K_i} (\vec{F}(\vec{w}) \cdot \vec{\phi}) dxdt.$$

The following fluxes can be introduced on the element boundaries

$$\hat{\vec{w}} = \vec{\lambda},$$
$$\hat{\mathbb{Z}} = \mathbb{Z} + (\mathbb{B}_{diff}(\vec{w} - \vec{\lambda})) \otimes \vec{n}.$$

where $\mathbb{B}_{diff}$ is a local stabilization matrix. The first flux is natural, as we are taking the trace values of $\vec{w}$ to be the Lagrange multipliers. The second flux can be traced back to the general form of the flux for the DG methods.

It is noticed that we have introduced additional degrees of freedom in the weak formulation (the $\lambda$'s) and thus in order to close the system, we need to add an additional equation. This equation is the so-called transmission condition [18] and is obtained by imposing the continuity of the normal flux (both advective and diffusive) weakly. This is due to that fact that all the spaces introduced so far have no continuity between the elements, however the conforming underlying finite element spaces of the continuous setting requires at least a minimum level of continuity across the elements. The transmission

condition would satisfy this requirement weakly and can be written as:

$$\int_{I_m} \sum_{i \in I} \int_{\partial K_i} \Big( \sum_{s=1}^{d} \vec{f}_s(\vec{\lambda}) n_s + \mathbb{B}_{ad}(\vec{w} - \vec{\lambda}) + \mathbb{Z} \cdot \vec{n} + \mathbb{B}_{diff}(\vec{w} - \vec{\lambda}) \Big) \cdot \vec{\mu} \, dsdt$$
$$= \int_{I_m} \sum_{i \in I} \int_{\partial K_i \cap \partial \Omega_N} \vec{g}_N \cdot \vec{\mu} dsdt \;\; \forall \vec{\mu} \in (H^k(I_m; H^p(\Gamma_{h,m})))^l(0) \tag{3.9}$$

where we have used (3.6) for the hybridized numerical flux. Note that the first term in equation (3.9) would cancel out on the interior edges, as the test function $\vec{\mu}$ is single-valued too. In case of using (3.7), this term needs to be computed. The two matrices $\mathbb{B}_{ad}$ and $\mathbb{B}_{diff}$ are called the local stabilization matrices and are crucial for stability, well-posedness and correct rate of convergence of the method.

Combining all the terms obtained so far, we get the following integral equations for the weak form:

Find $\vec{w} \in (H^k(I_m; H^p(\mathcal{T}_{h,m})))^l$, $\vec{\lambda} \in (H^k(I_m; H^p(\Gamma_{h,m})))^l(g_D)$ and $\mathbb{Z} \in$

$(H^k(I_m; H^p(\mathfrak{T}_{h,m})))^{l\times l}$ such that

$$\int_{I_m}(\sum_{i\in I}\int_{K_i}(\frac{\partial \vec{w}}{\partial t}\cdot\vec{\phi})dxdt+\sum_{i\in I}\int_{K_i}(\vec{w}(.,t_{m-1}^{+})-\vec{w}^{\uparrow})\cdot\vec{\phi}dx$$
$$+\int_{I_m}(\sum_{i\in I}\int_{\partial K_i\setminus\partial\Omega_D}\sum_{s=1}^{d}\vec{H}(\vec{w},\vec{\lambda},n_s)\cdot\vec{\phi}n_s)dsdt$$
$$+\int_{I_m}(\sum_{i\in I}\int_{\partial K_i\cap\partial\Omega_D}\sum_{s=1}^{d}\vec{H}(\vec{w},\vec{g}_D,n_s)\cdot\vec{\phi}n_s)dsdt-\int_{I_m}(\sum_{i\in I}\int_{K_i}\sum_{s=1}^{d}\vec{f}_s\cdot\frac{\partial\vec{\phi}}{\partial x_s})dxdt$$
$$-\int_{I_m}(\sum_{i\in I}\int_{K_i}\mathbb{Z}:\nabla\otimes\vec{\phi}dx)dt+\int_{I_m}(\sum_{i\in I}\int_{\partial K_i\setminus\Omega}(\mathbb{Z}\cdot\vec{n}+\mathbb{B}_{diff}(\vec{w}-\vec{\lambda}))\cdot\vec{\phi}dx)dt$$
$$+\int_{I_m}(\sum_{i\in I}\int_{\partial K_i\cap\partial\Omega_D}(\mathbb{Z}\cdot\vec{n}+\mathbb{B}_{diff}(\vec{w}-\vec{g_D}))\cdot\vec{\phi}dx)dt$$
$$+\int_{I_m}(\sum_{i\in I}\int_{\partial K_i\cap\partial\Omega_N}(\vec{g}_N+\mathbb{B}_{diff}(\vec{w}-\vec{\lambda}))\cdot\vec{\phi}dx)dt$$
$$=\int_{I_m}(\sum_{i\in I}\int_{K_i}\vec{F}(\vec{w})\cdot\vec{\phi}dx)dt,\quad\forall\vec{\phi}\in(H^k(I_m;H^p(\mathfrak{T}_{h,m})))^l,$$

$$\int_{I_m}(\sum_{i\in I}\int_{K_i}\mathbb{K}^{-1}\mathbb{Z}:\mathbb{V}dx)dt+\int_{I_m}(\sum_{i\in I}\int_{\partial K_i\setminus\partial\Omega_D}\vec{\lambda}\otimes\vec{n}:\mathbb{V}dx)dt$$
$$+\int_{I_m}(\sum_{i\in I}\int_{\partial K_i\cap\Omega_D}\vec{g}_D\otimes\vec{n}:\mathbb{V}dx)dt$$
$$-\int_{I_m}(\sum_{i\in I}\int_{K_i}\vec{w}\cdot(\nabla\cdot\mathbb{V}dx)dt=0,\quad\forall\,\mathbb{V}\in(H^k(I_m;H^p(\mathfrak{T}_{h,m})))^{l\times l},$$

$$\int_{I_m}\sum_{i\in I}\int_{\partial K_i}\big(\sum_{s=1}^{d}\vec{f}_s n_s+\mathbb{B}_{ad}(\vec{w}-\vec{\lambda})+\mathbb{Z}\cdot\vec{n}+\mathbb{B}_{diff}(\vec{w}-\vec{\lambda})\big)\cdot\vec{\mu}\,dsdt$$
$$=\int_{I_m}\sum_{i\in I}\int_{\partial K_i\cap\partial\Omega_N}\vec{g}_N\cdot\vec{\mu}dsdt\ \ \forall\vec{\mu}\in(H^k(I_m;H^p(\Gamma_{h,m})))^l(0)\qquad(3.10)$$

In the finite dimensional case, we define the finite element solution of the weak form (3.10) as:

**Definition 3.2.1.** We say that $\vec{w}_h \in (S^{p,q}_{h,\tau})^l, \vec{\lambda}_h \in (M^{p,q}_{h,\tau})^l(g_D), \mathbb{Z}_h \in (S^{p,q}_{h,\tau})^{l\times l}$ are the finite element solution of (3.10) if

$$\int_{I_m}(\sum_{i\in I}\int_{K_i}(\frac{\partial \vec{w}_h}{\partial t}\cdot\vec{\phi}_h)dxdt+\sum_{i\in I}\int_{K_i}(\vec{w}_h(.,t^+_{m-1})-\vec{w}^{\uparrow}_h)\cdot\vec{\phi}_h dx+$$

$$\int_{I_m}(\sum_{i\in I}\int_{\partial K_i\setminus\partial\Omega_D}\sum_{s=1}^{d}\vec{H}(\vec{w}_h,\vec{\lambda}_h,n_s)\cdot\vec{\phi}_h n_s)dsdt$$

$$+\int_{I_m}(\sum_{i\in I}\int_{\partial K_i\cap\partial\Omega_D}\sum_{s=1}^{d}\vec{H}(\vec{w}_h,P\vec{g}_D,n_s)\cdot\vec{\phi}_h n_s)dsdt$$

$$-\int_{I_m}(\sum_{i\in I}\int_{K_i}\sum_{s=1}^{d}\vec{f}_s\cdot\frac{\partial\vec{\phi}_h}{\partial x_s})dxdt.-\int_{I_m}(\sum_{i\in I}\int_{K_i}\mathbb{Z}_h:\nabla\otimes\vec{\phi}_h dx)dt$$

$$+\int_{I_m}(\sum_{i\in I}\int_{\partial K_i\setminus\Omega}(\mathbb{Z}_h\cdot\vec{n}+\mathbb{B}_{diff}(\vec{w}_h-\vec{\lambda}_h))\cdot\vec{\phi}_h dx)dt$$

$$+\int_{I_m}(\sum_{i\in I}\int_{\partial K_i\cap\partial\Omega_D}(\mathbb{Z}_h\cdot\vec{n}+\mathbb{B}_{diff}(\vec{w}_h-P\vec{g}_D))\cdot\vec{\phi}_h dx)dt$$

$$+\int_{I_m}(\sum_{i\in I}\int_{\partial K_i\cap\partial\Omega_N}(P\vec{g}_N\cdot\vec{n}+\mathbb{B}_{diff}(\vec{w}_h-\vec{\lambda}))\cdot\vec{\phi}_h dx)dt$$

$$=\int_{I_m}(\sum_{i\in I}\int_{K_i}\vec{F}(\vec{w}_h)\cdot\vec{\phi}_h dx)dt,\quad \forall\vec{\phi}_h\in(S^{p,q}_{h,\tau})^l,$$

$$\int_{I_m}(\sum_{i\in I}\int_{K_i}\mathbb{K}^{-1}\mathbb{Z}_h:\mathbb{V}_h dx)dt+\int_{I_m}(\sum_{i\in I}\int_{\partial K_i\setminus\partial\Omega_D}\vec{\lambda}_h\otimes\vec{n}:\mathbb{V}_h dx)dt$$

$$+\int_{I_m}(\sum_{i\in I}\int_{\partial K_i\cap\Omega_D}P\vec{g}_D\otimes\vec{n}:\mathbb{V}_h dx)dt$$

$$-\int_{I_m}(\sum_{i\in I}\int_{K_i}\vec{w}_h\cdot(\nabla\cdot\mathbb{V}_h dx)dt=0,\quad\forall\,\mathbb{V}_h\in(S^{p,q}_{h,\tau})^{l\times l},$$

$$\int_{I_m} \sum_{i\in I} \int_{\partial K_i} \big( \sum_{s=1}^{d} \vec{f}_s n_s + \mathbb{B}_{ad}(\vec{w}_h - \vec{\lambda}_h) + \mathbb{Z}_h \cdot \vec{n} + \mathbb{B}_{diff}(\vec{w}_h - \vec{\lambda}_h) \big) \cdot \vec{\mu}_h \, dsdt$$
$$= \int_{I_m} \sum_{i\in I} \int_{\partial K_i \cap \partial\Omega_N} P\vec{g}_N \cdot \vec{\mu}_h dsdt \;\; \forall \vec{\mu}_h \in (M_{h,\tau}^{p,q}(0))^l. \tag{3.11}$$

# Chapter 4

# STHDG for Shallow Water Equations

In this chapter we follow the same approach developed in the previous chapter to obtain the variational formulation for shallow water equations. Then we cover the numerical implementation of the method used. Later we consider the specific treatment of the source term in shallow water equations and explain the method used for obtaining a well-balanced scheme.

## 4.1 Weak formulation

Consider the following system of shallow water equations obtained in Chapter 2 written in a compact form

$$\begin{aligned} &\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \vec{u}) = 0 \\ &\frac{\partial (\rho \vec{u})}{\partial t} + \nabla \cdot (\rho \vec{u} \otimes \vec{u}) + g \nabla \frac{\rho^2}{2} = -g\rho(\nabla b)^T \end{aligned} \tag{4.1}$$

where we are using $\rho$ for water height instead of $h$ and $\vec{u}$ is the velocity of the fluid. We would now like to extend the (4.1) to the so called viscous shallow water equations. The correct mathematical form of viscosity to be added is [64]

$$-\mathbb{D}\nabla \cdot (\rho \nabla \vec{u}),$$

where $\mathbb{D}$ is the viscosity tensor. This is a non-conservative product term and as was explained in Chapter 2 we are trying not to deal with non-conservative terms as the mathematical definition of these terms are not well-defined. Therefore we are going to adopt a simplification and modify the term by

$$-\mathbb{D}\nabla \cdot \nabla(\rho\vec{u}).$$

This form of diffusion has been used before in the literature [1]. In our case the rationale behind that is that eventually we are going to use a small difussivity coefficient and thus the diffusion term would not play a major role in our computations. We choose

$$\mathbb{D} = \begin{bmatrix} \epsilon_x & 0 \\ 0 & \epsilon_y \end{bmatrix},$$

where $\epsilon_x$ and $\epsilon_x$ are diffusion coefficients in $x$ and $y$ direction, respectively. Therefore considering a Dirichlet boundary conditions, the final form of the equations to be investigated will be

$$\begin{aligned}
&\frac{\partial \rho}{\partial t} + \frac{\partial q_x}{\partial x} + \frac{\partial q_y}{\partial y} = 0 \\
&\frac{\partial q_x}{\partial t} + \frac{\partial}{\partial x}(\frac{q_x^2}{\rho} + \frac{g\rho^2}{2}) + \frac{\partial}{\partial y}(\frac{q_x q_y}{\rho}) - \epsilon_x \Delta q_x = -g\rho b_x \\
&\frac{\partial q_y}{\partial t} + \frac{\partial}{\partial x}(\frac{q_x q_y}{\rho}) + \frac{\partial}{\partial y}(\frac{q_y^2}{\rho} + \frac{g\rho^2}{2}) - \epsilon_y \Delta q_y = -g\rho b_y \\
&\rho|_{t=0} = \rho_0, \quad q_x|_{t=0} = q_{x0}, \quad q_y|_{t=0} = q_{y0} \\
&\rho = \rho^b, \quad q_x = q_x^b, \quad q_y = q_y^b \quad on \quad \partial\Omega \times (0,T).
\end{aligned} \tag{4.2}$$

where $q_x$ and $q_y$ are fluxes in the $x$ and $y$ direction respectively and $b$ is the bottom topography. Note that as stated in Chapter 2, the number of boundary

conditions needed is a function of the flow regime. However to be definite we have chosen to include the Dirichlet boundary condition to be imposed for all the variables. Before writing the weak form, we introduce the following notations,

$$
\begin{aligned}
(u,v)_K &= \int_K uvdx, \\
(u,v)_{\mathcal{T}_{h,m}} &= \sum_{K_i \in \mathcal{T}_{h,m}} \int_{K_i} uvdx \\
(u,v)_{I_m \times K} &= \int_{I_m} \int_K uvdxdt, \\
(u,v)_{I_m \times \mathcal{T}_{h,m}} &= \sum_{K_i \in \mathcal{T}_{h,m}} \int_{I_m} \int_{K_i} uvdxdt \\
\langle u,v \rangle_{I_m \times \Gamma_{h,m}} &= \sum_{K_i \in \mathcal{T}_{h,m}} \int_{I_m} \int_{\partial K_i} uvdsdt
\end{aligned}
$$

The meaning of other similar notations may be inferred from these.

We are using upwinding in time, hybridization in space, Lax-Friedrichs flux for advective flux and mixed hybrid formulation for diffusion. Following the same procedure as in the previous chapter we obtain the following weak formulation:

**Definition 4.1.1.** We say that $\rho, q_x, q_y, \lambda_\rho, \lambda_{q_x}, \lambda_{q_y}, \sigma_x, \sigma_y \in S^{p,q}_{h,\tau} \times S^{p,q}_{h,\tau} \times S^{p,q}_{h,\tau} \times M^{p,q}_{h,\tau}(g_\rho) \times M^{p,q}_{h,\tau}(g_{q_x}) \times M^{p,q}_{h,\tau}(g_{q_y}) \times (S^{p,q}_{h,\tau})^2 \times (S^{p,q}_{h,\tau})^2$ are the STHDG finite element solution of the system (4.2) if

continuity: $\forall \varphi_1 \in S_{h,\tau}^{p,q}$

$$(\partial_t \rho, \varphi_1)_{I_m \times \mathcal{T}_{h,m}} + (\rho(., t_{m-1}^+) - \rho^{\uparrow}(., t_{m-1}^-), \varphi_1)_{\mathcal{T}_{h,m}} + \langle \lambda_{q_x}, \varphi_1 n_x \rangle_{I_m \times \Gamma_{h,m}^{int}}$$
$$+ \langle g_{q_x}, \varphi_1 n_x \rangle_{I_m \times \Gamma_{h,m}^{b}} - (q_x, \partial_x \varphi_1)_{I_m \times \mathcal{T}_{h,m}} + \langle \lambda_{q_y}, \varphi_1 n_y \rangle_{I_m \times \Gamma_{h,m}^{int}} + \langle g_{q_y}, \varphi_1 n_y \rangle_{I_m \times \Gamma_{h,m}^{b}}$$
$$- (q_y, \partial_y \varphi_1)_{I_m \times \mathcal{T}_{h,m}} + \alpha_{c_\rho} \langle (\rho - \lambda_\rho), \varphi_1 \rangle_{I_m \times \Gamma_{h,m}^{int}} + \alpha_{c_\rho} \langle (\rho - g_\rho), \varphi_1 \rangle_{I_m \times \Gamma_{h,m}^{b}} = 0,$$

x-momentum: $\forall \varphi_2 \in S_{h,\tau}^{p,q}$

$$(\partial_t q_x, \varphi_2)_{I_m \times \mathcal{T}_{h,m}} + (q_x(., t_{m-1}^+) - {q_x}^{\uparrow}(., t_{m-1}^-), \varphi_2)_{\mathcal{T}_{h,m}} + \langle \frac{\lambda_{q_x}^2}{\lambda_\rho} + g\frac{\lambda_\rho^2}{2}, \varphi_2 n_x \rangle_{I_m \times \Gamma_{h,m}^{int}}$$
$$+ \langle \frac{g_{q_x}^2}{g_\rho} + g\frac{g_\rho^2}{2}, \varphi_2 n_x \rangle_{I_m \times \Gamma_{h,m}^{b}} - (\frac{{q_x}^2}{\rho} + \frac{g\rho^2}{2}, \partial_x \varphi_2)_{I_m \times \mathcal{T}_{h,m}} + \langle \frac{\lambda_{q_y}\lambda_{q_x}}{\lambda_\rho}, \varphi_2 n_y \rangle_{I_m \times \Gamma_{h,m}^{int}}$$
$$+ \langle \frac{g_{q_y} g_{q_x}}{g_\rho}, \varphi_2 n_y \rangle_{I_m \times \Gamma_{h,m}^{b}} - (\frac{q_x q_y}{\rho}, \partial_y \varphi_2)_{I_m \times \mathcal{T}_{h,m}} + (\alpha_{cq_x} + \alpha_{dq_x}) \langle (q_x - \lambda_{q_x}), \varphi_2 \rangle_{I_m \times \Gamma_{h,m}^{int}}$$
$$+ (\alpha_{cq_x} + \alpha_{dq_x}) \langle (q_x - g_{q_x}), \varphi_2 \rangle_{I_m \times \Gamma_{h,m}^{b}} + \langle \vec{\sigma}_x \cdot \vec{n}, \varphi_2 \rangle_{I_m \times \Gamma_{h,m}} - (\vec{\sigma}_x, \nabla \varphi_2)_{I_m \times \mathcal{T}_{h,m}} = 0,$$

y-momentum: $\forall \varphi_3 \in S_{h,\tau}^{p,q}$

$$(\partial_t q_y, \varphi_3)_{I_m \times \mathcal{T}_{h,m}} + (q_y - {q_y}^{\uparrow}, \varphi_3)_{\mathcal{T}_{h,m}} + \langle \frac{\lambda_{q_y}^2}{\lambda_\rho} + g\frac{\lambda_\rho^2}{2}, \varphi_3 n_y \rangle_{I_m \times \Gamma_{h,m}^{int}}$$
$$+ \langle \frac{g_{q_y}^2}{g_h} + g\frac{g_\rho^2}{2}, \varphi_3 n_y \rangle_{I_m \times \Gamma_{h,m}^{b}} - (\frac{{q_y}^2}{\rho} + \frac{g\rho^2}{2}, \partial_y \varphi_3)_{I_m \times \mathcal{T}_{h,m}} + \langle \frac{\lambda_{q_y}\lambda_{q_x}}{\lambda_\rho}, \varphi_3 n_x \rangle_{I_m \times \Gamma_{h,m}^{int}}$$
$$+ \langle \frac{g_{q_y} g_{q_x}}{g_\rho}, \varphi_3 n_x \rangle_{I_m \times \Gamma_{h,m}^{b}} - (\frac{q_x q_y}{\rho}, \partial_x \varphi_3)_{I_m \times \mathcal{T}_{h,m}} + (\alpha_{cq_y} + \alpha_{dq_y}) \langle (q_y - \lambda_{q_y}), \varphi_3 \rangle_{I_m \times \Gamma_{h,m}^{int}}$$
$$+ (\alpha_{cq_y} + \alpha_{dq_y}) \langle (q_y - g_{q_y}), \varphi_3 \rangle_{I_m \times \Gamma_{h,m}^{b}} + \langle \vec{\sigma}_y \cdot \vec{n}, \varphi_3 \rangle_{I_m \times \Gamma_{h,m}} - (\vec{\sigma}_y, \nabla \varphi_3)_{I_m \times \mathcal{T}_{h,m}} = 0,$$

diffusive fluxes: $\forall \varphi_i \in S_{h,\tau}^{p,q}, \quad i \in \{4,5,6,7\}$

$$\frac{1}{\epsilon_1}(\sigma_{x_1}, \varphi_4)_{I_m \times \mathcal{T}_{h,m}} + \langle \lambda_{q_x}, \varphi_4 n_x \rangle_{I_m \times \Gamma_{h,m}^{int}} + \langle g_{q_x}, \varphi_4 n_x \rangle_{I_m \times \Gamma_{h,m}^{b}} - (q_x, \partial_x \varphi_4)_{I_m \times \mathcal{T}_{h,m}} = 0,$$
$$\frac{1}{\epsilon_1}(\sigma_{x_2}, \varphi_5)_{I_m \times \mathcal{T}_{h,m}} + \langle \lambda_{q_x}, \varphi_5 n_y \rangle_{I_m \times \Gamma_{h,m}^{int}} + \langle g_{q_x}, \varphi_5 n_y \rangle_{I_m \times \Gamma_{h,m}^{b}} - (q_x, \partial_y \varphi_5)_{I_m \times \mathcal{T}_{h,m}} = 0,$$
$$\frac{1}{\epsilon_2}(\sigma_{y_1}, \varphi_6)_{I_m \times \mathcal{T}_{h,m}} + \langle \lambda_{q_y}, \varphi_6 n_x \rangle_{I_m \times \Gamma_{h,m}^{int}} + \langle g_{q_y}, \varphi_6 n_x \rangle_{I_m \times \Gamma_{h,m}^{b}} - (q_y, \partial_x \varphi_6)_{I_m \times \mathcal{T}_{h,m}} = 0,$$
$$\frac{1}{\epsilon_2}(\sigma_{y_2}, \varphi_7)_{I_m \times \mathcal{T}_{h,m}} + \langle \lambda_{q_y}, \varphi_7 n_y \rangle_{I_m \times \Gamma_{h,m}^{int}} + \langle g_{q_y}, \varphi_7 n_y \rangle_{I_m \times \Gamma_{h,m}^{b}} - (q_y, \partial_y \varphi_7)_{I_m \times \mathcal{T}_{h,m}} = 0,$$

transmission conditions:

$$\langle \alpha_{c_\rho}(\rho - \lambda_\rho), \mu_1 \rangle_{I_m \times \Gamma_{h,m}} = 0, \qquad \forall \mu_1 \in M^{p,q}_{h,\tau}(0)$$
$$\langle (\alpha_{cq_x} + \alpha_{dq_x})(q_x - \lambda_{q_x}) + \vec{\sigma}_x \cdot \vec{n}, \mu_2 \rangle_{I_m \times \Gamma_{h,m}} = 0, \quad \forall \mu_2 \in M^{p,q}_{h,\tau}(0)$$
$$\langle (\alpha_{cq_y} + \alpha_{dq_y})(q_y - \lambda_{q_y}) + \vec{\sigma}_y \cdot \vec{n}, \mu_3 \rangle_{I_m \times \Gamma_{h,m}} = 0, \quad \forall \mu_3 \in M^{p,q}_{h,\tau}(0)$$

where $\vec{\sigma}_x = (\sigma_{x_1}, \sigma_{x_2}) = -\epsilon_1 \nabla q_x$, $\vec{\sigma}_y = (\sigma_{y_1}, \sigma_{y_2}) = -\epsilon_2 \nabla q_y$ and the $\alpha$'s are the local stabilization parameters with the subindices "$c$" and "$d$" corresponding to convection and diffusion, respectively.

As stated, the values of diffusion coefficients are taken to be small, i.e. $(\epsilon_x = \epsilon_y = 1 \times 10^{-12}\, \frac{\mathrm{m}^2}{\mathrm{s}})$ in the simulations, therefore the values of $\alpha$'s corresponding to diffusion are set to zero. Regarding the advection, based on the analysis done in Chapter 2, these values are taken to be the greatest eigenvalue of the system, i.e.

$$\alpha_{c_\rho} = \alpha_{c_{qx}} = \alpha_{c_{qy}} = |\frac{q_x}{\rho} n_x + \frac{q_y}{\rho} n_y| + \sqrt{g\rho},$$
$$\alpha_{d_{qx}} = \alpha_{d_{qy}} \approx 0.$$

The weak form obtained is a system of ten equations in ten unknowns. We need to linearize these equations before solving them. We choose the Newton method for linearization.

## 4.2 Linearization

We use the Taylor expansion for linearization of the equations, thus for a given function $f(p)$, we substitute it as $p = \bar{p} + \delta p$ and apply the Taylor

expansion, i.e.

$$f(\bar{p}+\delta p) \approx f(\bar{p}) + (\frac{\partial f}{\partial p})|_{\bar{p}} \cdot \delta p$$

As an example the linearized form of the continuity equation can be written as:

$$\begin{aligned}
&(\partial_t \delta\rho, \varphi_1)_{I_m\times\mathcal{T}_{h,m}} + (\delta\rho, \varphi_1)_{\mathcal{T}_{h,m}} + \langle \delta\lambda_{q_x}, \varphi_1 n_x\rangle_{I_m\times\Gamma^{int}_{h,m}} - (\delta q_x, \partial_x\varphi_1)_{I_m\times\mathcal{T}_{h,m}} \\
&+ \langle \delta\lambda_{q_y}, \varphi_1 n_y\rangle_{I_m\times\Gamma^{int}_{h,m}} - (\delta q_y, \partial_y\varphi_1)_{I_m\times\mathcal{T}_{h,m}} + \alpha_{c_\rho}\langle \delta\rho, \varphi_1\rangle_{I_m\times\Gamma_{h,m}} \\
&- \alpha_{c_\rho}\langle \delta\lambda_\rho, \varphi_1\rangle_{I_m\times\Gamma^{int}_{h,m}} = -\alpha_{c_\rho}\langle \bar{\rho}, \varphi_1\rangle_{I_m\times\Gamma_{h,m}} + \alpha_{c_\rho}\langle \bar{\lambda}_\rho, \varphi_1\rangle_{I_m\times\Gamma^{int}_{h,m}} \\
&+ \alpha_{c_\rho}\langle g_\rho, \varphi_1\rangle_{I_m\times\Gamma^{b}_{h,m}} - (\partial_t\bar{\rho}, \varphi_1)_{I_m\times\mathcal{T}_{h,m}} - \langle \bar{\lambda}_{q_x}, \varphi_1 n_x\rangle_{I_m\times\Gamma^{int}_{h,m}} + (\bar{q_x}, \partial_x\varphi_1)_{I_m\times\mathcal{T}_{h,m}} \\
&- \langle \bar{\lambda}_{q_y}, \varphi_1 n_y\rangle_{I_m\times\Gamma^{int}_{h,m}} + (\bar{q_y}, \partial_y\varphi_1)_{I_m\times\mathcal{T}_{h,m}} + (\rho^{\uparrow}, \varphi_1)_{\mathcal{T}_{h,m}} - \langle g_{q_x}, \varphi_1 n_x\rangle_{I_m\times\Gamma^{b}_{h,m}} \\
&- \langle g_{q_y}, \varphi_1 n_y\rangle_{I_m\times\Gamma^{b}_{h,m}}, \qquad (4.3)
\end{aligned}$$

and the same for other equations obtained in the weak form. By assembling all the matrices of the above system, we end up having the following form:

$$\begin{pmatrix} A & B & C \\ D & E & F \\ G & H & I \end{pmatrix} \begin{pmatrix} \delta U \\ \delta Q \\ \delta\Lambda \end{pmatrix} = \begin{pmatrix} J \\ K \\ L \end{pmatrix} \qquad (4.4)$$

where $U = \{\rho, q_x, q_y\}^T$, $Q = \{\sigma_{x_1}, \sigma_{x_2}, \sigma_{y_1}, \sigma_{y_2}\}^T$, $\Lambda = \{\lambda_\rho, \lambda_{q_x}, \lambda_{q_y}\}^T$ and the submatrices $A, B, ..., I$ are the corresponding terms obtained in the weak form. We can now solve for $U$ and $Q$ in terms of $\Lambda$ and obtain

$$\begin{pmatrix} \delta U \\ \delta Q \end{pmatrix} = \begin{pmatrix} A & B \\ D & E \end{pmatrix}^{-1} \left\{ \begin{pmatrix} J \\ K \end{pmatrix} - \begin{pmatrix} C \\ F \end{pmatrix} \delta\Lambda \right\} \qquad (4.5)$$

As the underlying finite element spaces are in $L^2$ (not continuous across each element) each submatrices $A, B, D, E$ is block diagonal and thus the inversion

can be done locally at each element level. Expanding the last row of the matrix in (4.4) and using (4.5) we obtain

$$\mathbb{K}\Lambda = \mathbb{F} \tag{4.6}$$

where

$$\mathbb{K} = -\begin{pmatrix} G & H \end{pmatrix} \begin{pmatrix} A & B \\ D & E \end{pmatrix}^{-1} \begin{pmatrix} C \\ F \end{pmatrix} + I$$

and

$$\mathbb{F} = L - \begin{pmatrix} G & H \end{pmatrix} \begin{pmatrix} A & B \\ D & E \end{pmatrix}^{-1} \begin{pmatrix} J \\ K \end{pmatrix}.$$

The equations (4.6) are the Schur complement of the system (4.4). It is the discretized version of the Poincare-Steklov operator, operating between the trace spaces $\Lambda$ and its dual $\Lambda'$.

## 4.3 Well-balanced formulation

The presence of source terms in hyperbolic equations modifies their analytical properties in comparison with the homogeneous case. More specifically, they lead to various steady state solutions to be satisfied, resulting from the balance between source terms and internal forces. In the case of shallow water equations with topography, one of these steady states is the "lake at rest" condition for which the fluxes are zero and $\rho + b = constant$. The difficulty is to preserve the steady state solutions at the discrete level. Schemes which can preserve the unperturbed steady state at the discrete level are called well-balanced schemes. Initially for scalar problems, Greenberg, LeRoux and

others introduced the notion of well-balanced schemes (see [15],[16] for details). This definition has been further developed by Gosse and LeRoux [11], which used a reformulation of the source terms by means of non-conservative products to derive numerical fluxes at the interfaces of an unstructured mesh. An approach by Leveque [27] is based on the Godunov scheme extended for an appropriately modified system. Botchorishvili, Perthame and Vasseur presented in [2] a kinetic scheme, that maintains steady states and which is proved to converge when stiff source terms are considered. Using interfacial values, instead of the cell-averages, for the source term, Jin proposed in [17] a rather simple method for capturing steady state solutions with a high order accuracy. Previous schemes have also been modified for this target by Bermudez and Vasquez [6]. These kinds of numerical processing have been extended to hyperbolic systems of balance laws (like the Saint-Venant system for shallow waters), to obtain stable schemes which preserve the steady states (see e.g. [35] [39] [37]).

Our formulation is based on the idea of hydrostatic reconstruction [4] modified for hybridization. First note that in the finite dimensional level, the equality $\rho_h + b = constant$ can only be satisfied when $b$ is in the same space as $\rho$ which can be seen by moving $\rho_h$ to the right hand side. So the first step is to $L^2$ project the bathymetry into the space of $\rho_h$. The basic idea in hydrostatic reconstruction is to evaluate the convective fluxes at hydrostatic reconstructed water height and its modified flux. In order to illustrate, we consider the momentum and mass equations in (4.2) separately . First consider

the momentum equations for the steady state situation, (i.e $\rho = \rho(x)$ and $q_x = q_y = 0$), multiplied by test function and integrated over an element,

$$\int_K \partial_x \begin{pmatrix} \frac{g\rho^2}{2} \\ 0 \end{pmatrix} \cdot \phi + \int_K \partial_y \begin{pmatrix} 0 \\ \frac{g\rho^2}{2} \end{pmatrix} \cdot \phi = -\int_K \begin{pmatrix} g\rho b_x \\ g\rho b_y \end{pmatrix} \cdot \phi$$

Note that equilibrium is satisfied, as in the steady state, $b_x = -\rho_x, b_y = -\rho_y$. Performing an integration by parts forward and then backward, we end up having

$$\int_K \partial_x \begin{pmatrix} \frac{g\rho^2}{2} \\ 0 \end{pmatrix} \cdot \phi + \int_K \partial_y \begin{pmatrix} 0 \\ \frac{g\rho^2}{2} \end{pmatrix} \cdot \phi + \int_{\partial K} \begin{pmatrix} (\frac{g\lambda_\rho^2 - g\rho^2}{2}) n_x \\ (\frac{g\lambda_\rho^2 - g\rho^2}{2}) n_y \end{pmatrix} \cdot \phi = -\int_K \begin{pmatrix} g\rho b_x \\ g\rho b_y \end{pmatrix} \cdot \phi$$

Note that the third integral on the left will not make the equation well-balanced. Therefore if this term is simply added to right hand side we can obtain equilibrium for the momentum equations at the discrete level. This same idea was originally used in the hydrostatic reconstruction. As there is no single-valued trace in the case of DG or Finite Volume method, a reconstructed single-valued water height was proposed as

$$\rho_h^{rec} = (\rho_h + b_h - max(b_h|_{K^+}, b_h|_{K^-}))_+$$

where $x_+ = max(0, x)$. And then the fluxes were modified accordingly

$$q_x^{rec} = \frac{q_x \times \rho_h^{rec}}{\rho_h}, \quad q_y^{rec} = \frac{q_y \times \rho_h^{rec}}{\rho_h}.$$

In our case the value of $\lambda_\rho$ at which fluxes are evaluated is between the left and right water height. This can be shown from the first transmission condition as follows. In the case that the trace value of $\rho$ is in the same space as $\lambda_\rho$, the equation is satisfied pointwise and for an interior edge it can be written as

$$\alpha_{c\rho}^+(\rho^+ - \lambda_\rho) + \alpha_{c\rho}^-(\rho^- - \lambda_\rho) = 0$$

Solving for $\lambda_\rho$ we obtain

$$\lambda_\rho = \frac{\alpha^+_{c\rho}}{\alpha^+_{c\rho} + \alpha^-_{c\rho}}\rho^+ + \frac{\alpha^-_{c\rho}}{\alpha^+_{c\rho} + \alpha^-_{c\rho}}\rho^- \tag{4.7}$$

which is a convex combination of the left and right water height and therefore lies between the two. This is preferable as the $\lambda_\rho$ remains positive as long as the water height is.

The modification of the flux would change the transmission conditions corresponding to momentum equations in the weak form as the hydrostatic flux $\frac{q\lambda^2_\rho}{2}$ is replaced with $\frac{g\rho^2}{2}$. Therefore it must be included in the transmission conditions.

Now considering the equation of conservation of mass, from the weak form (4.3) it can be seen that there is a residual due to Lax-Friedrichs flux at steady state, for an element this is equal to

$$\alpha^+_{c\rho}(\rho^+ - \lambda_\rho), \tag{4.8}$$

which is not zero at steady state. The difference between the water height at left and right states is the difference between their corresponding bathymetry. Without loss of generality, let assume $\rho^- = \rho^+ - \Delta b$. Inserting this value in equation (4.7) and then insert the result back in (4.8), we obtain

$$\alpha^+_{c\rho}(\rho^+ - \lambda_\rho) = \alpha^+_{c\rho}(\rho^+ - \rho^+ + \frac{\alpha^-_{c\rho}}{\alpha^+_{c\rho} + \alpha^-_{c\rho}}\Delta b) = \frac{\alpha^-_{c\rho}\alpha^+_{c\rho}}{\alpha^+_{c\rho} + \alpha^-_{c\rho}}\Delta b.$$

This term can be added to the right hand side of mass conservation to get a well-balanced scheme. In the case where $\alpha^+_{c\rho} = \alpha^-_{c\rho} = \alpha$ this term is equal to $\frac{\alpha}{2}\Delta b$.

## 4.4 Stabilization of shock

For shallow water equations as a set of hyperbolic equations, shocks may develop in the domain. These shocks create undershoots and overshoots due to Gibbs phenomena and thus local extrema can be created. This is not in agreement with the structure of the underlying PDE which usually satisfy a version of maximum principle nor with the numerical scheme which generally needs to be total variation bounded (TVB) or total variation diminishing (TVD). Therefore shocks need to be stabilized. There are essentially two ways to stabilize the shock, either by adding extra diffusion terms to the weak form or by post processing techniques often called slope limiters. Each of the methods has its own advantages and disadvantages. A disadvantage of the first method is that the imposed diffusion would affect the areas of smooth solution too which is not desirable, which can be by-passed in slope-limiter case by using shock detector. On the other hand slope limiters, especially in the case of explicit methods, are expensive as they need to be applied at each iteration resulting in an increase in computational time. Slope limiters are used in the simulations in this thesis.

### 4.4.1 Shock detector

In order to apply the limiter only when necessary, a shock detector needs to be used. For shock detection, the criterion proposed in [46] is used

which can be written as [31]

$$\mathcal{I}_K := \sum_{e^- \in \partial K} \frac{|\int_{\partial e^-} (\rho^+ - \rho^-)|}{h_K^{(p+1)\backslash 2} |e^-| \, |\langle \rho \rangle_K|}$$

where $e^-$ corresponds to the inflow edge of an element and $\langle \rho \rangle_K$ denotes the mean value of water height. The limiting should be applied when $\mathcal{I}_K \geq 1$, otherwise there is no need to apply the limiter to the solution.

### 4.4.2 Slope limiter

As we are using space-time (as an implicit) method, the limiter is only applied at the end of time step. Different types of slope limiter were used both for rectangular and triangular prism elements. The original slope limiter by Cockburn and Shu [21], the one by Hoteit et al [42] in which the limited solution is obtained by minimization of a least square problem and the vertex-based limiter of Barth and Jespersen [5] were implemented. For each simulation the type of limiter used will be mentioned.

## 4.5 Implementation

Both 1D and 2D shallow water equations are implemented based on STHDG method. The weak form for 1D was not presented in this chapter and it can be obtained following the same approach explained for 2D case. For 1D shallow water (2D code), quad space-time elements are used. For 2D shallow water (3D code), cube and prism elements are implemented. A schematic view of cube elements along with trace (quad) elements is shown in figure 4.1.

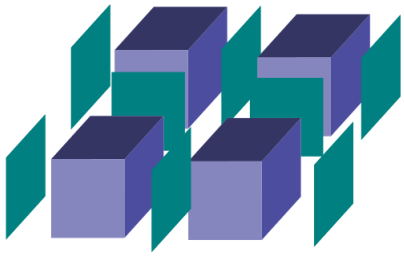

Figure 4.1: A schematic view of cube space-time elements and quadrilateral space-time trace elements used for implementation

The overall solution procedure is shown in Algorithm 1. The only detail that needs to be mentioned is that when the solution converges in the Newton iteration, the slope limiter is applied. This procedure would change the solution but the trace values are not updated accordingly. Thus another loop will be performed to update the trace values while the solution is fixed via imposing a constraint.

**Algorithm 1** Implementation of HDG

```
1: for t = 1...T do
2:     while .not.Converged do
3:         for i = 1 : ne do
4:             Calculate element matrices
5:             Calculate Schur complement (locally invert the element matrix)
6:             Assemble the condensed matrix on the global "trace" matrix
7:         end for
8:         Solve for traces
9:         for i = 1 : ne do
10:            Calculate element matrices
11:            Recover solution values back from traces
12:        end for
13:        Check for convergence
14:        if Converged & .not.Limited then
15:            Apply slope limiter
16:            Converged = False (Put constraint on limited solution, loop to
    update trace values)
17:        end if
18:    end while
19: end for
```

# Chapter 5

# Numerical Results

In this chapter, we investigate the effectiveness of our methodology as applied to a collection of shallow water benchmark problems: circular dam break problem, flow through constricted channel, partial dam break, tidal flow simulation of the Bahamas islands.

At first some preliminary results will be shown for the 1D shallow water case. In the following, the results for the 2D case will be discussed. In order to show the effectiveness of the method, most of the problems are run on coarse meshes, with no h/p adaptivity, no post-processing (except for limiter), fairly large time steps (compared to explicit DG method), linear space-time elements, and the most simplest limiter without shock detector (e.g. [5]). The diffusion coefficients are taken to be $\epsilon_x = \epsilon_y = 1 \times 10^{-12} \frac{\mathrm{m}^2}{\mathrm{s}}$.

## 5.1 1D shallow water

Before moving to the 2D shallow water case, the 1D model was simulated via the STHDG method. Square elements are used for space-time finite element and linear elements to model the trace values. The details of the weak formulation is not covered in this thesis and the reader may follow essentially

the same approach as in 2D case.

### 5.1.1 Dam break problem

The dam break problem is one of the classical verification tests for numerical discretization of SWE. Based on the initial water height we can have different flow regimes. With respect to numerical test, the super critical flow is a more challenging one which corresponds to higher water shock. The test is simulated with an initial shock of 8 m as shown in Figure 5.1. In order to show the inherent capabilities of the method without embellishing it, a coarse mesh with an element size of 8 m, a time slab of 0.1 s, Lax-Friedrichs flux and the simplest possible limiter, minmod limiter is used. Limiter is applied at the end of time step only. The water height as well the flux are shown in Figure 5.2 at time $t = 20$ s along with their corresponding analytical solution. The blue line is the FE solution, the blue circles are the trace values (Lagrange multipliers) and the red plot is the exact solution. As can be seen a fair agreement is obtained between the exact solution and the simulation results considering the coarse mesh, numerical flux and limiter used.

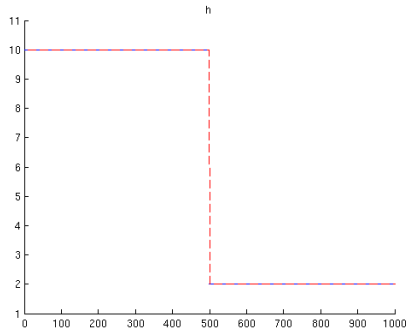


Figure 5.1: Water height ($\rho$) at t = 0.0 s

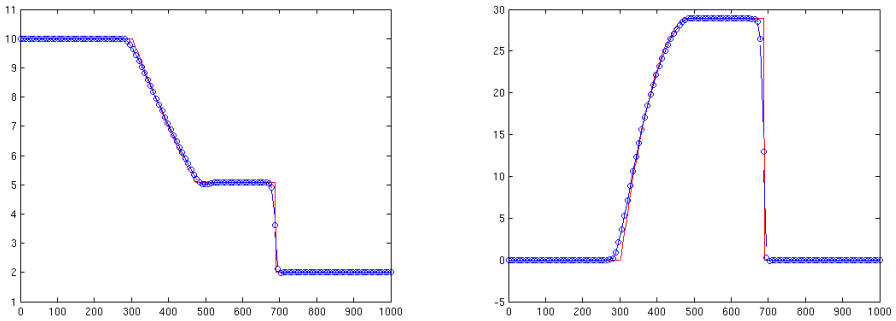


Figure 5.2: Water height $\rho$(m) (left) and Flux $q(\frac{\mathrm{m}^2}{\mathrm{s}})$ (right) at t = 20 s

## 5.2 2D shallow water

The weak formulation is explained in previous chapters. Both cube and prism space-time elements are implemented in the code. As triangular meshes are more used in practice for simulations, we will only cover triangular prism elements in the rest of this chapter. Trace elements are square. We are using Lax-Friedrichs for advective flux.

### 5.2.1 Circular dam break problem

For the first benchmark test, we are going to simulate the circular dam break problem. This test case consists of the instantaneous breaking of a cylindrical tank initially filled with water at rest. The wave generated by the breaking of the tank propagates into still water surrounding the tank. The test is useful to check the ability of the method to preserve the cylindrical symmetry. Indeed, the problem becomes 1D in the radial direction and the governing equations, rewritten with reference to a radial coordinate system, can be written as

$$\frac{\partial \rho}{\partial t} + \frac{\partial q_r}{\partial x} = -\frac{q_r}{r},$$
$$\frac{\partial q_r}{\partial t} + \frac{\partial}{\partial x}(\frac{q_r^2}{\rho} + g\frac{\rho^2}{2}) = -\frac{q_r^2}{r\rho},$$

where $r$ is the radius and $q_r = \rho u_r$, where $u_r$ is the velocity in the radial direction. We first need to verify the solution. We use the test in [10] for our simulation, where the diameter of the cylinder is 20 m, the initial water height inside the cylinder is 2 m and the surrounding water height is 0.5 m. We are using 13440 prism elements and the time step is 0.05 s. Figure 5.3 compares the solution along a given radial direction at times $t = 1$ s and $t = 2.5$ s with a solution given in [10]. There is a good agreement between the two plot specifically in capturing the critical points. No spurious oscillations are visible.

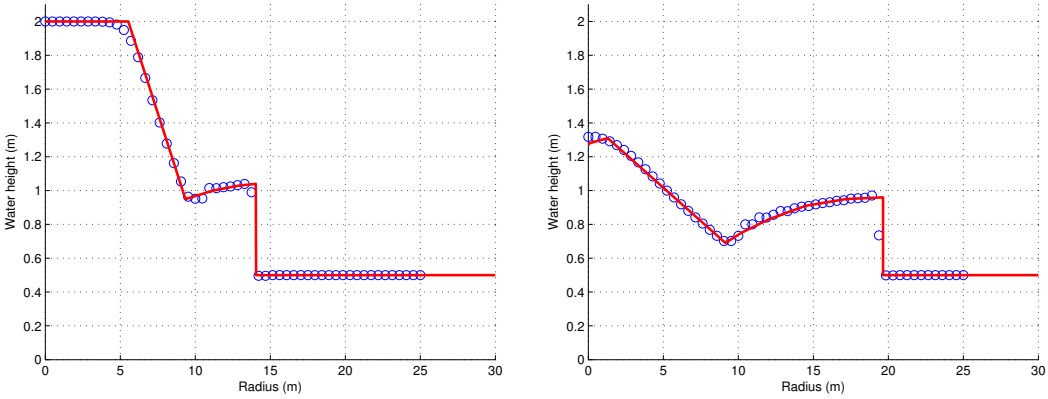


Figure 5.3: Water elevation at times $t = 1\,\mathrm{s}$ (left) and $t = 2.5\,\mathrm{s}$ (right). The blue circles are the STHDG solution and the red line is the 1D solution given in [10], computed with 10,000 cells and PRICE-C scheme

We now move to a more challenging problem of super-critical dam break flow .The diameter of the cylinder is 22 m, the initial water height inside the cylinder is 10 m and the surrounding water height is 1.0 m (a 9 m shock, see Figure 5.4) . We are using 4880 prism elements and the time step is 0.05 s. For limiter we are using that of Barth and Jespersen [5]. The water heights at times $t = 0.4\,\mathrm{s}, t = 0.8\,\mathrm{s}$ and $t = 1.8\,\mathrm{s}$ are shown in the Figures 5.4-5.7

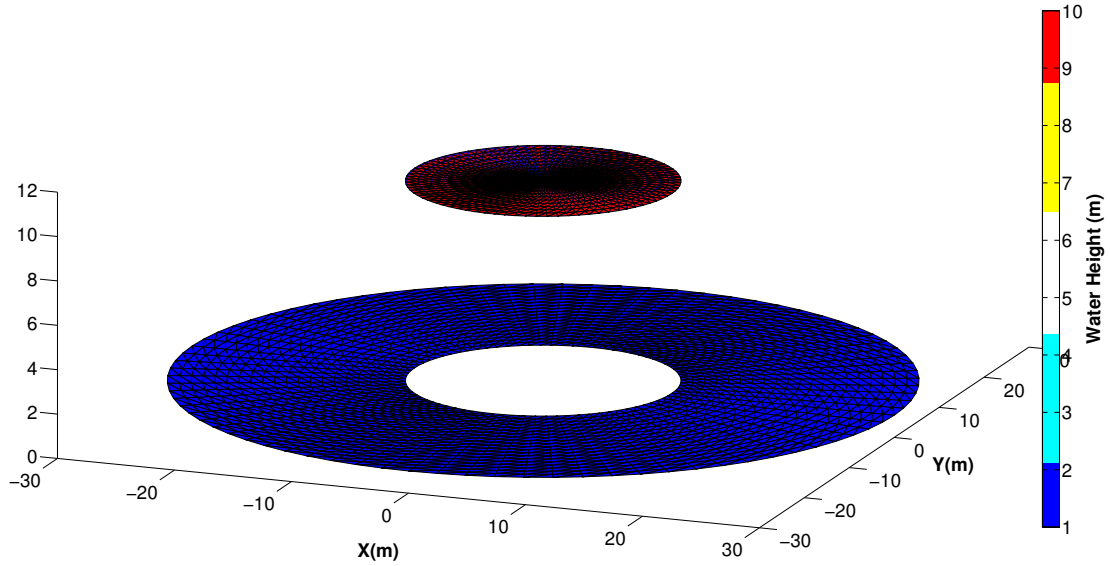


Figure 5.4: Water height ($\rho$) at $t = 0.0\,\mathrm{s}$

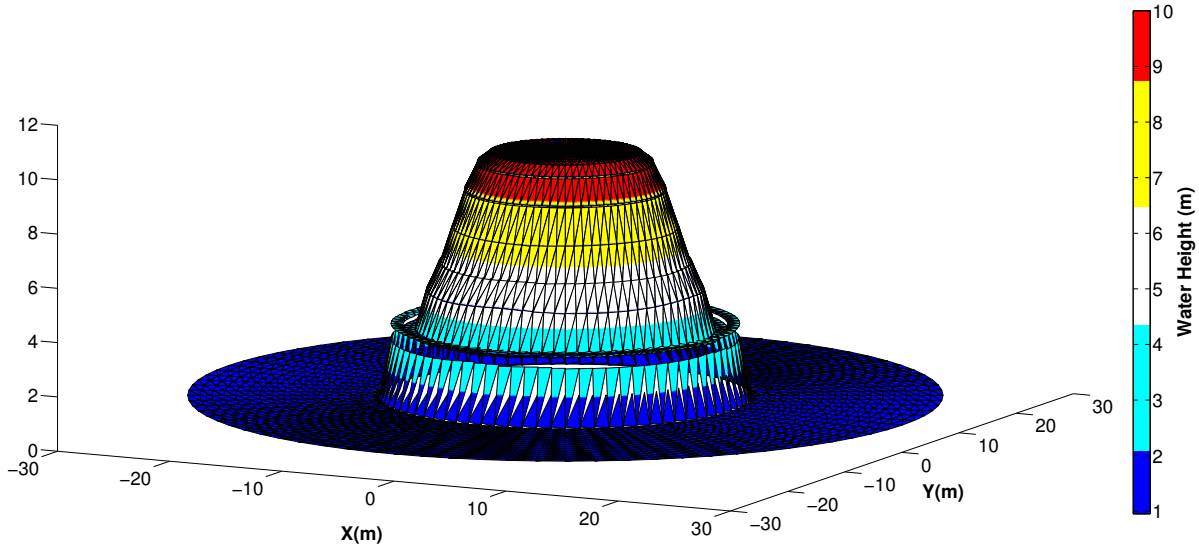


Figure 5.5: Water height ($\rho$) at $t = 0.4\,\mathrm{s}$

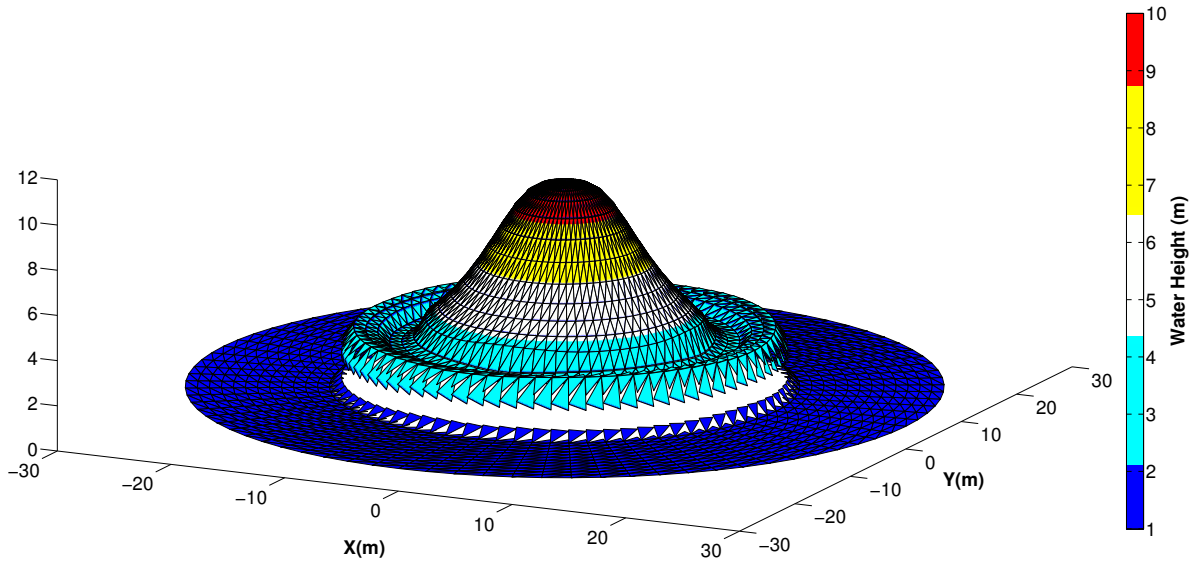


Figure 5.6: Water height ($\rho$) at $t = 0.8\,\mathrm{s}$

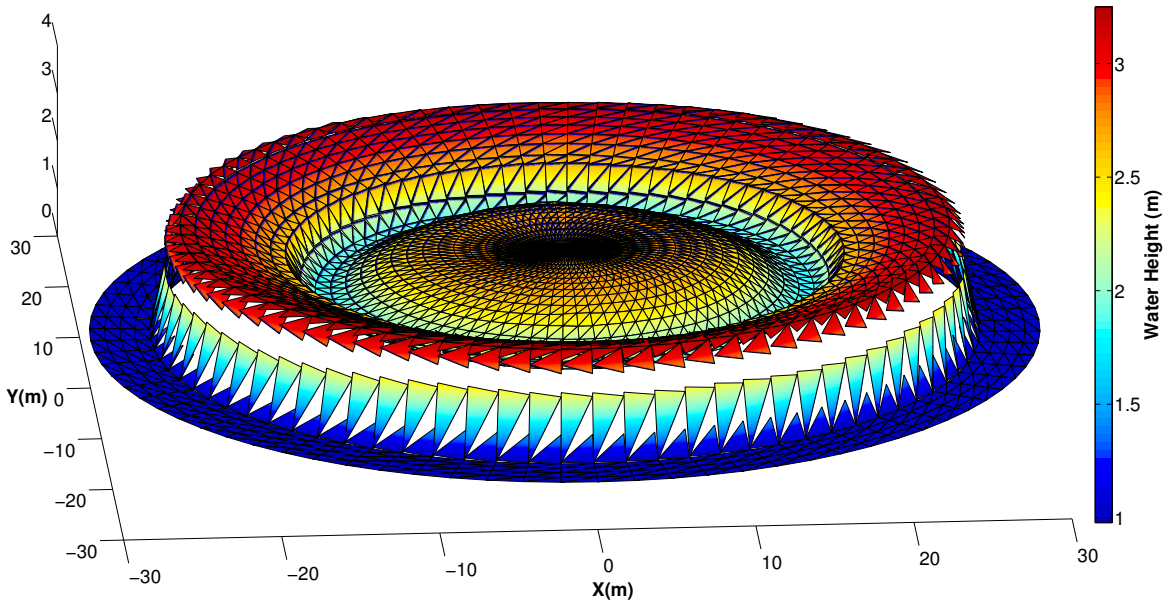


Figure 5.7: Water height ($\rho$) at $t = 1.8\,\mathrm{s}$

As can be seen no over shoot or under shoot is observed near the shocks. An excellent symmetry is observed in the results.

### 5.2.2 Supercritical flow through a contraction

Supercritical channel flows subject to a change in the cross-section can lead to the formation of shock and rarefaction waves. Here, we take the configuration as in [63]. The constricted angle is $\alpha = 5^\circ$. We consider a flat bed b = 0 and initial condition $\rho = 1m$, $q_x = \sqrt{g\rho}F$, $q_y = 0$, $g = 9.81$ and Froude number $F = 2.5$ (for $F > 1$ the flow is super critical). The inflow boundary condition on the left is the same as initial condition and we are using an outflow boundary condition on the right. This simply means that value of the Lagrange multipliers on the boundary would be equal to the corresponding face attached to it (extrapolation from inside). The top and bottom boundary conditions are zero normal flux i.e. $q \cdot n = 0$. This can be imposed by modifying the fluxes as

$$\hat{q}_x = q_x - (q \cdot n)n_x$$

$$\hat{q}_y = q_y - (q \cdot n)n_y.$$

We are using 8120 prism space-time element. One time step of $\Delta t = 2.0\,\mathrm{s}$ is chosen. The numerical values of water height are shown in in Figures 5.8 (plan view) and 5.9 (3D view) at time $t = 20\,\mathrm{s}$. The analytical values of the two plateaus formed are $h = 1.25\,\mathrm{m}$ and $h = 1.55\,\mathrm{m}$ (see [50]) which are in an excellent agreement with the results obtained. The $q_x$ and $q_y$ fluxes are shown in Figures 5.10 and 5.11 respectively.

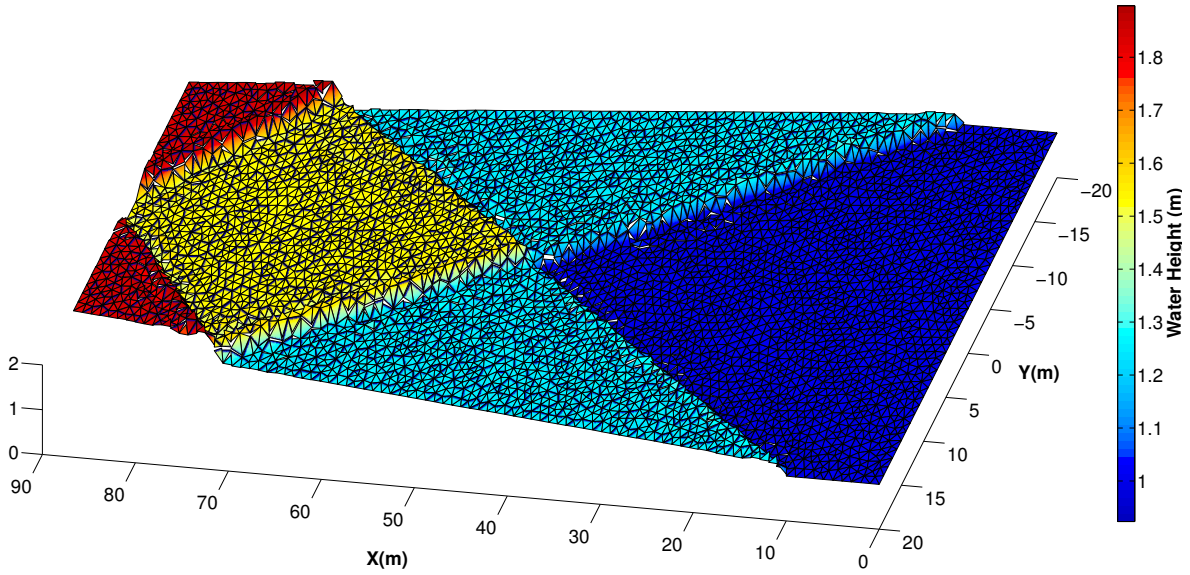


Figure 5.8: Water height $\rho$ at $t = 20\,\mathrm{s}$ (3D view)

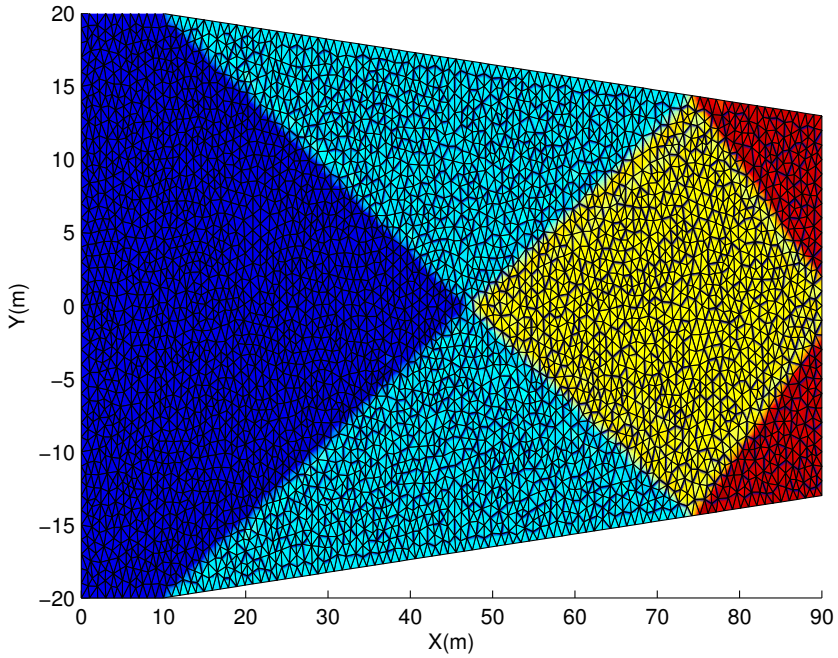


Figure 5.9: Water height ($\rho$) at $t = 20\,\mathrm{s}$ (plan view)

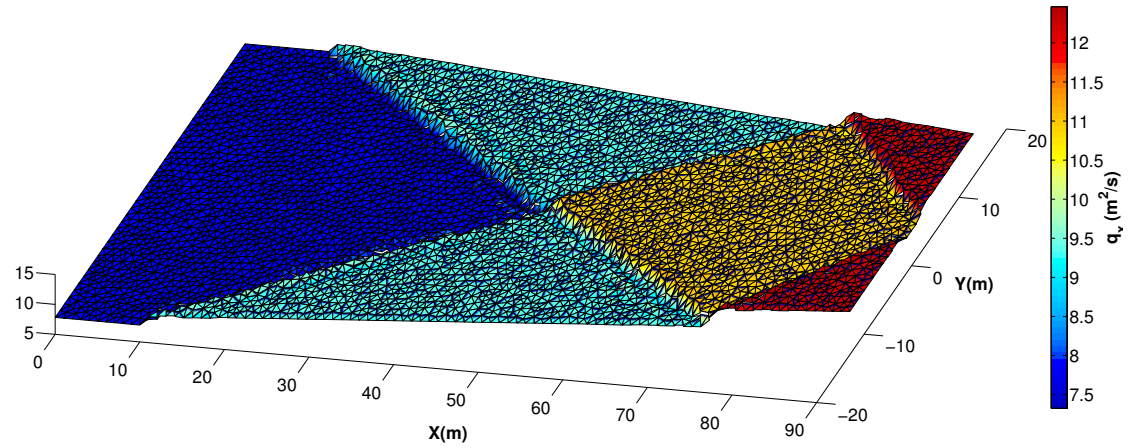


Figure 5.10: Flux in $x$ direction ($q_x$) at $t = 20\,\mathrm{s}$

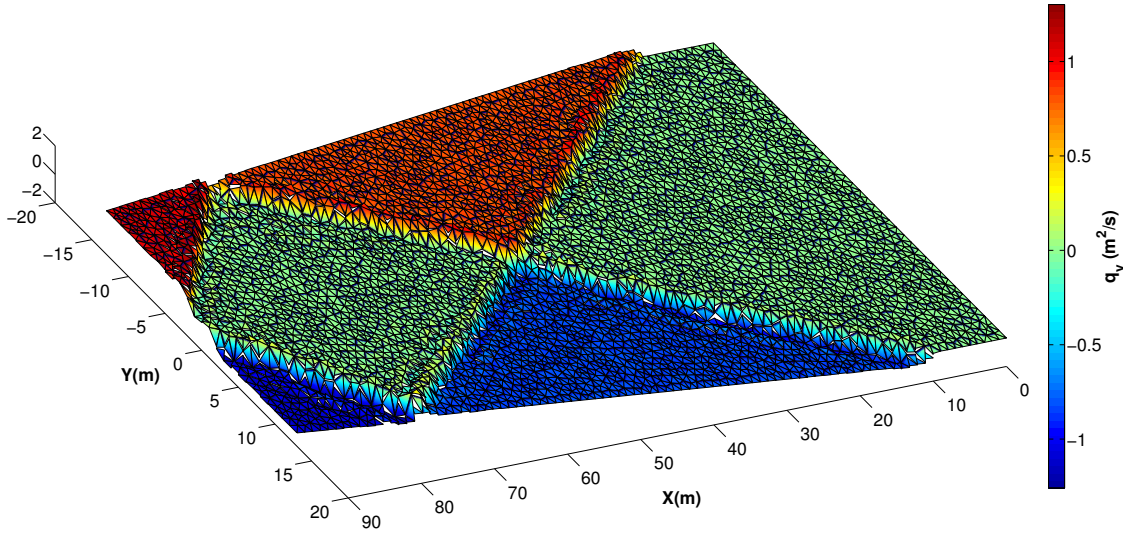


Figure 5.11: Flux in $y$ direction ($q_y$) at $t = 20\,\mathrm{s}$

### 5.2.3 Partial dam break

Partial dam break is another benchmark problem. The initial setup is shown in Figure 5.12, where the water height is 10 m and 5 m at upstream and downstream respectively. The time step is $\Delta t = 0.25$ s and 3648 elements are used. The results of simulation after 5 s of simulation are shown in Figures 5.13-5.15. The water height at the breach is 7.5 m which is the same as the one reported in [69]. When there is a finite water depth downstream, a shock front always exists. The results can be compared with those in [69]. There is a very good agreement in numerical values, however our model better captures the shock and the details of the flow.

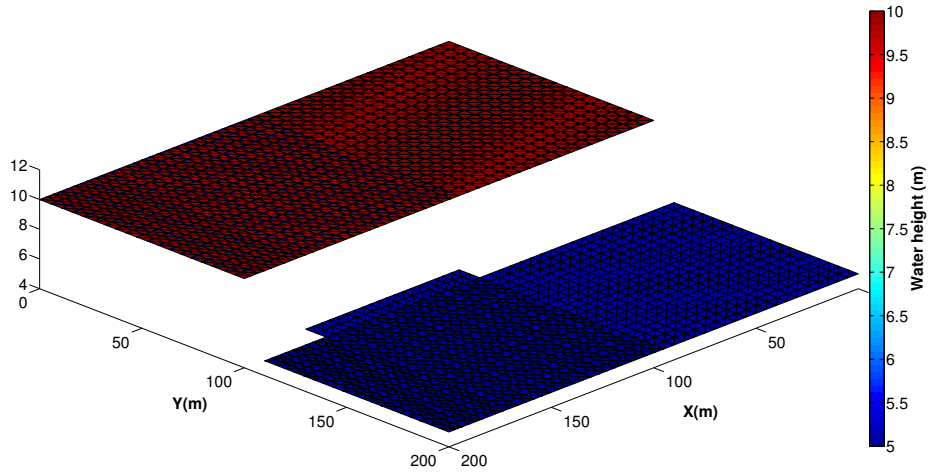


Figure 5.12: Initial water height

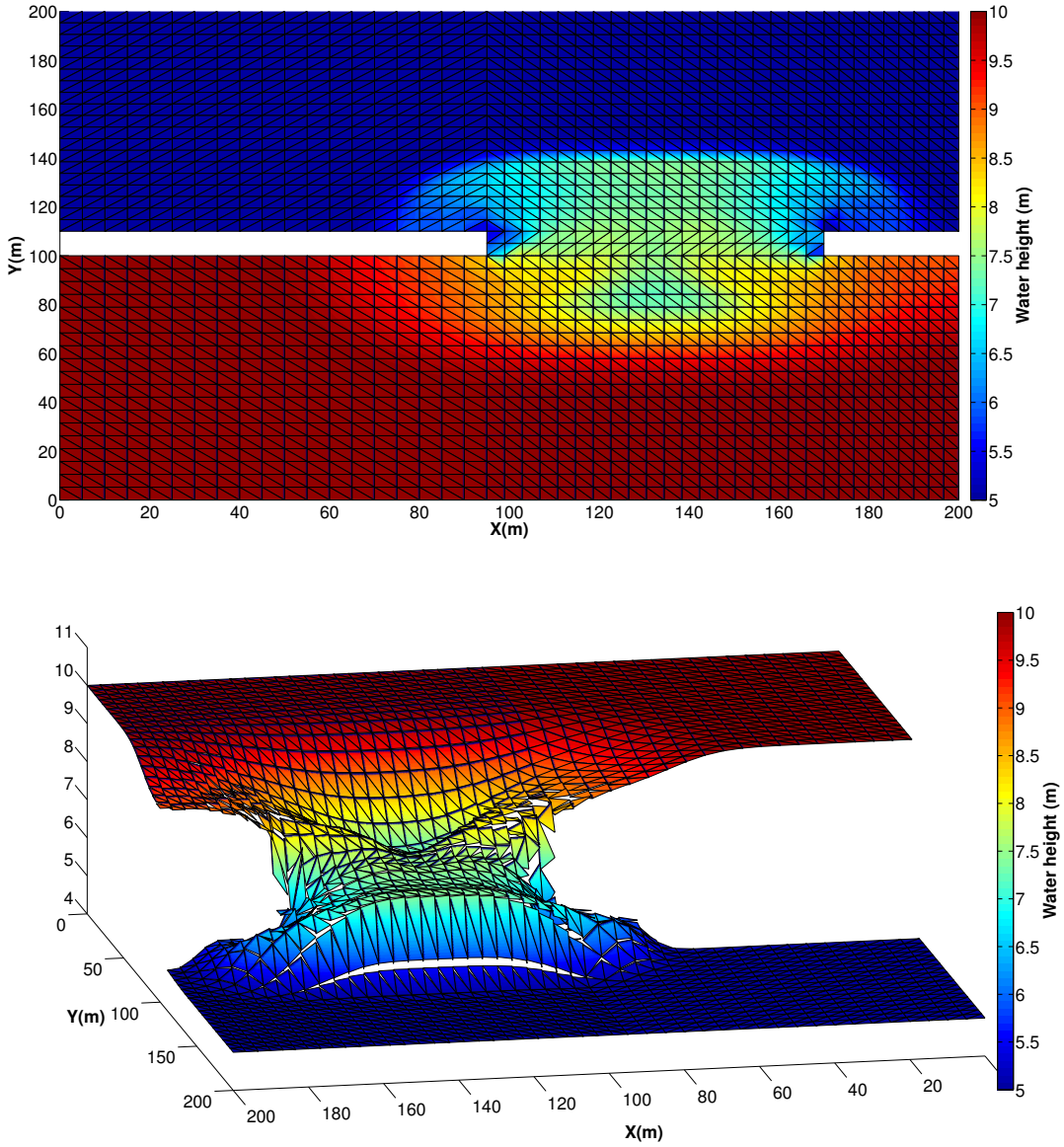


Figure 5.13: Water height $\rho$ at time $t = 5.0\,\mathrm{s}$

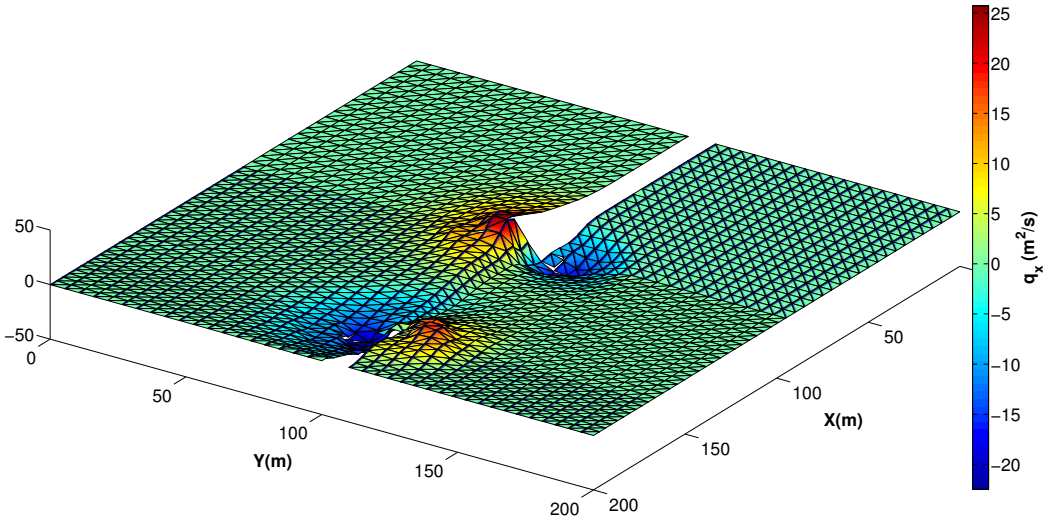


Figure 5.14: Flux $q_x$ at time $t = 5.0\,\mathrm{s}$

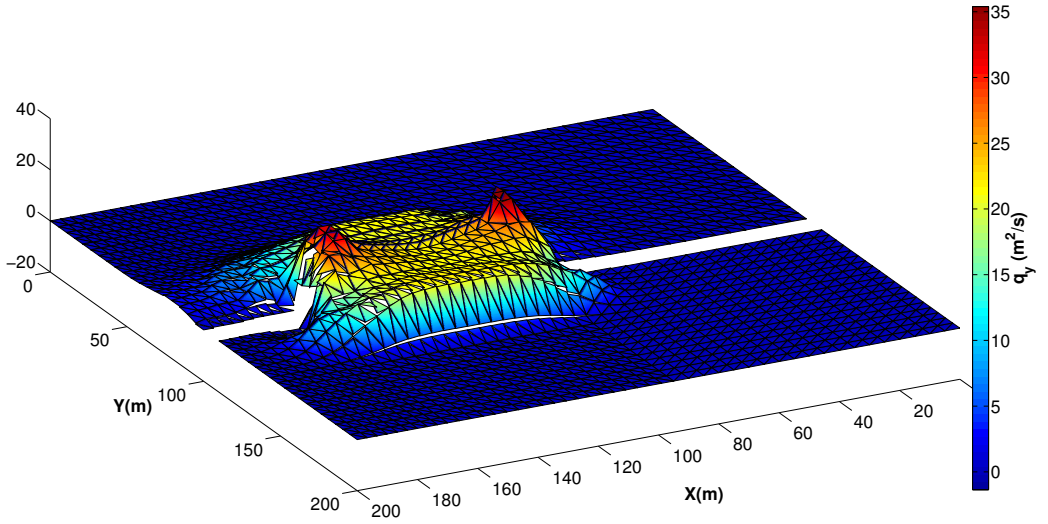


Figure 5.15: Flux $q_y$ at time $t = 5.0\,\mathrm{s}$

### 5.2.4 Well-balanced test

The purpose of the first test problem is to verify the well-balanced property of our algorithm towards the steady-state solution. We consider the test case in [66]. The computational domain is a square of size $[0, 1] \times [0, 1]$. The bottom bathymetry function is chosen as:

$$b(x, y) = max(0, 1 - (10x - 5)^2 - (10y - 5)^2)$$

and the initial data is taken as

$$\rho(x, y, 0) = 2 - b(x, y), \quad q_x(x, y, 0) = 0, \quad q_y(x, y, 0) = 0.$$

This steady state should be exactly preserved, and the surface should remain flat. We compute the solution until t = 0.5 s on the triangular meshes with the mesh size 0.1 m. In order to demonstrate that the still water solution is indeed maintained up to round-off error, we use the double-precision to perform the computation, and show the $L^1$ error for the water height $\rho$ and the discharges $q_x$ and $q_y$ in Table 5.1. We can see that all errors are at the level of round-off errors, which verifies the well-balanced property. Also the water height along with the topography is shown Figure 5.16.

| $L^1$ error | | |
|---|---|---|
| $\rho$ | $q_x$ | $q_y$ |
| $6.7752e-14$ | $1.1495e-12$ | $1.2689e-12$ |

Table 5.1: $L^1$ error for the stationary solution at $t = 0.5$ s

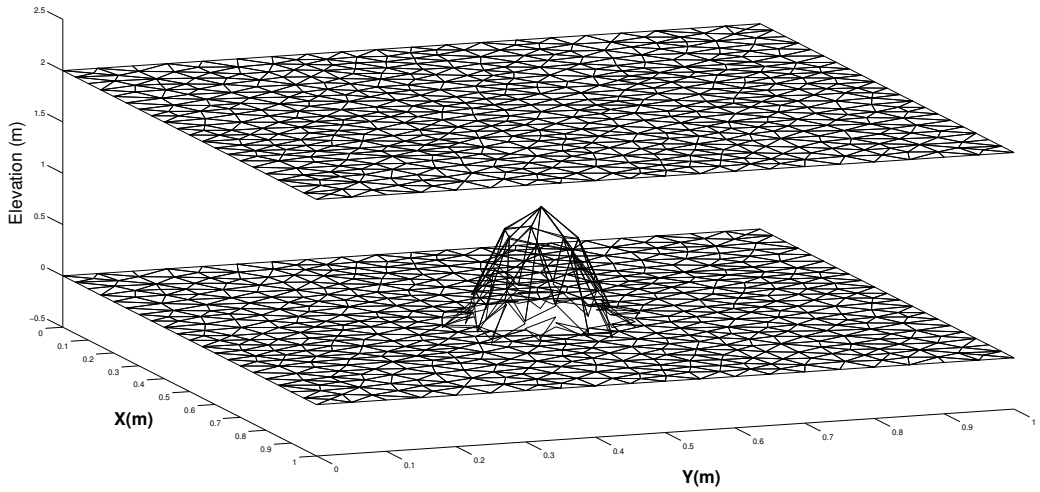


Figure 5.16: Bottom topography ($b$) and total water height ($\rho + b = 2.0\,\mathrm{m}$) at $t = 0.5\,\mathrm{s}$

### 5.2.5 The Bahamas Islands

As a final numerical example we consider a more practical problem of tide-driven flow near the Bahamas Islands. The bathymetry of the domain is shown in Figure 5.17. We are imposing an open sea boundary condition on the straight line on the right and a land boundary condition on the rest (for the details please refer to [1]). The following tidal forcing function was imposed

at the open sea boundary:

$$\begin{aligned}\xi(t) =& 0.075cos(\frac{t}{25.82} + 3.40) \\ &+0.095cos(\frac{t}{23.94} + 3.60) \\ &+0.1cos(\frac{t}{12.66} + 5.93) \\ &+0.395cos(\frac{t}{12.42} + 0.00) \\ &+0.06cos(\frac{t}{12.00} + 0.75)\end{aligned}$$

where time $t$ is in hours and $\xi$ in meters. The $\xi(0) \neq 0$ and therefore the time was shifted such that we have approximately $\xi(0) = 0$. This is equivalent to start imposing the boundary condition at $t \approx 18\,\mathrm{h}$.

We are tracking the water height at four different stations. The location of these stations are shown in Figure 5.18 whose coordinates in meters are (55166.672; 10166.665), (44250.000; 29333.335), (38666.664; 49333.328) and (24500.000; 89500.000). The simulation is cold-started. In contrast to reference [1], the tidal force is applied without any ramp-up. The ramp-up is usually a tangent hyperbolic function applied to the forcing function to gradually increase the forcing to its full values over several days and it's crucial for the stability of the code. Also no bottom friction is applied. Note that both of these factors are crucial to obtain a stable solution from a DG code. The authors tested the ADCIRC code by eliminating any of these two factors and the code essentially could not converge within the first several time steps.

The time step used was variable and we started with 1 h time step at the beginning, for up to a few hours. However we had issues with the convergence

of the Newton method and therefore we tried to both reducing the time step and using a damped version of the Newton method. We chose an adaptive strategy by reducing the time steps to 0.5 h, 20 min and 10 min if we had an issue with convergence. Also a damped version of Newton method was used adaptively, starting from full Newton method and then reducing the damped factor to that of $\alpha = 0.5$, $\alpha = 0.2$ and $\alpha = 0.1$.

The time history of the water elevation at different stations are plotted in Figures 5.20 and 5.21 for more than four days. Some important features of the plots are first the they follow the same pattern of the imposed boundary condition and second that the maximum and minimum of the water heights are approximately the same as those of $\xi$ which shows that the overall behavior is correct. We observe some oscillations in the numerical results of Stations 1-3, which is not present in station 4. This could be due to different factors such as the geometry of the island as we are essentially seeing this type of oscillation only on the left hand side of the island.

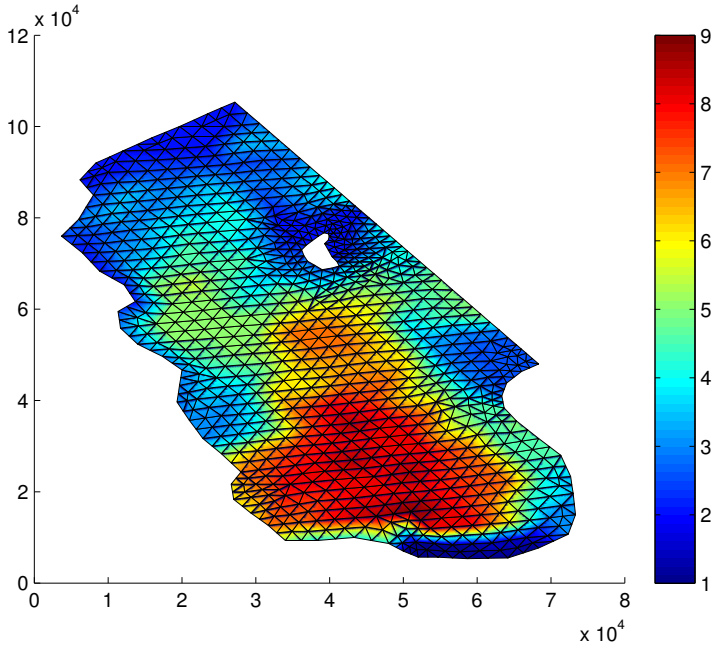


Figure 5.17: Bathymetry

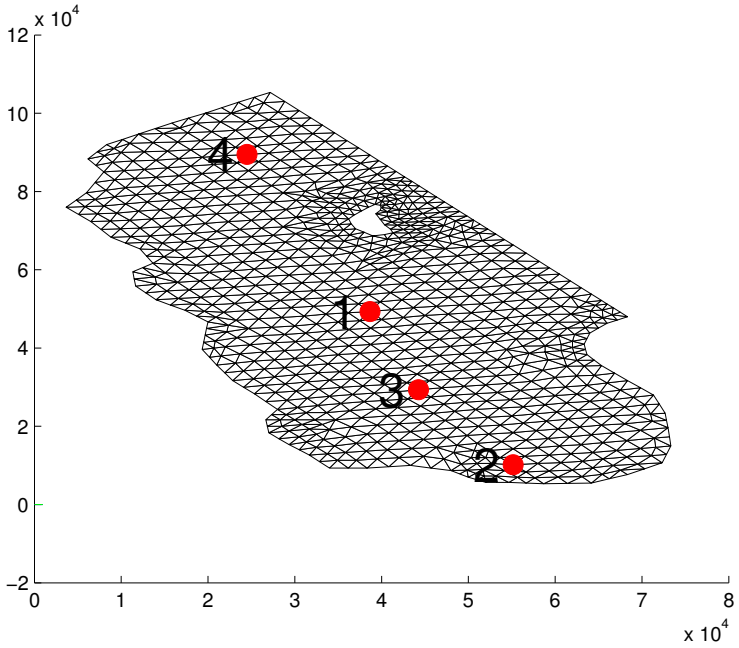


Figure 5.18: Stations $1 - 4$ accross the domain

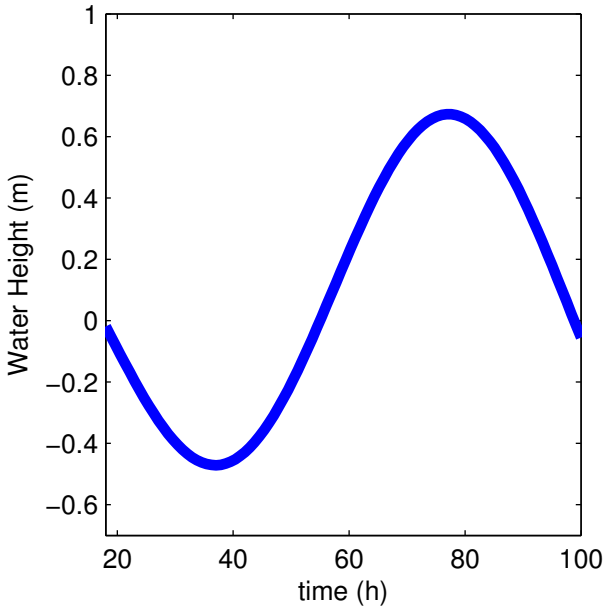


Figure 5.19: Time history of the applied boundary condition $\zeta(t)$

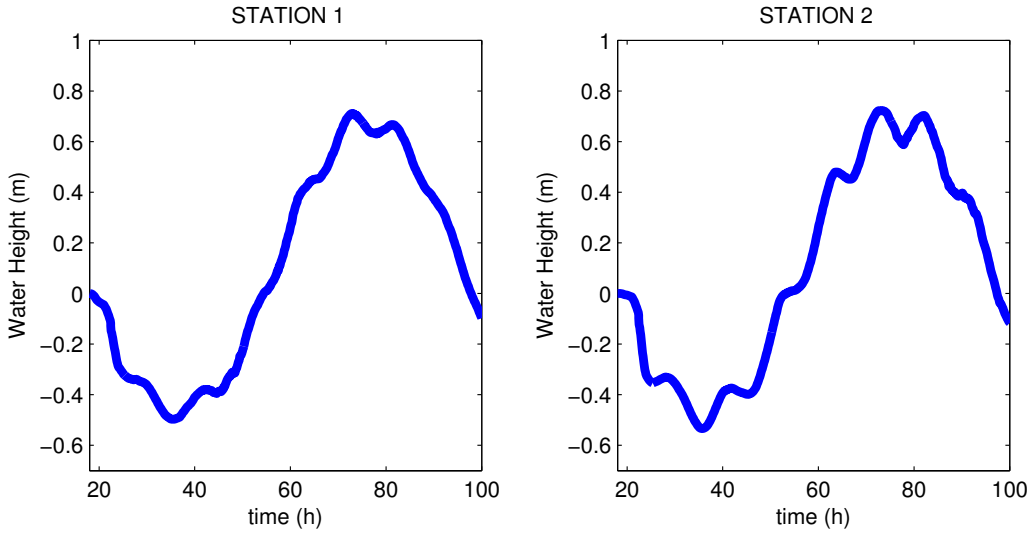


Figure 5.20: Time histories of water elevation (excluding bathymetry) at Stations 1-2

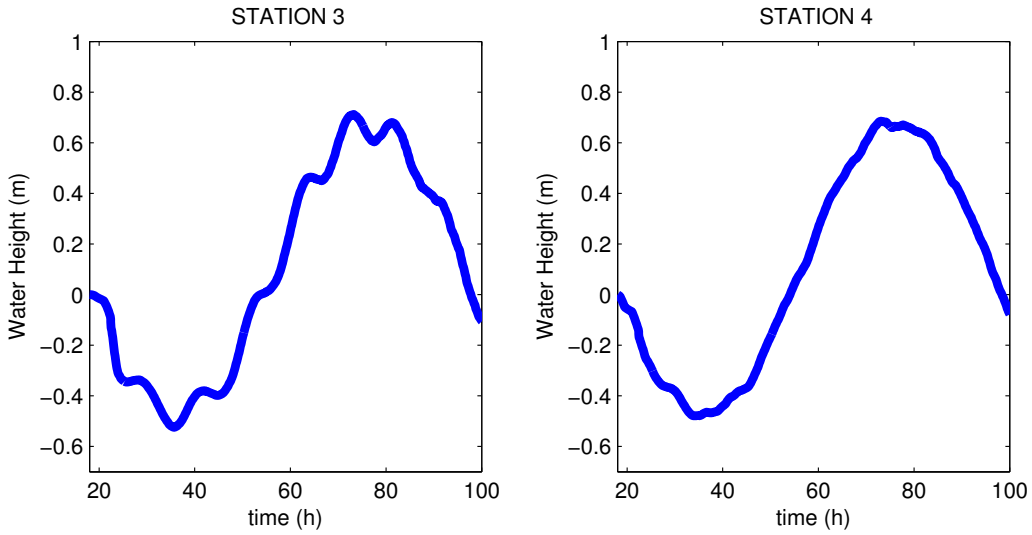


Figure 5.21: Time histories of water elevation (excluding bathymetry) at Stations 3-4

# Chapter 6

# An *a priori* error estimate

In this chapter an *a priori* error estimate will be derived for the shallow water equations via space-time HDG method. At first the main assumptions are stated, then the space-time projections used will be introduced. We will then derive an abstract error estimate in terms of the sizes of the space and time meshes.

## 6.1 Formulation of the problem

Consider the following set of shallow water equations

$$
\begin{aligned}
&\partial_t \rho + \partial_x q_x + \partial_y q_y = 0; \\
&\partial_t q_x + \partial_x (\frac{q_x^2}{\rho} + \frac{g\rho^2}{2}) + \partial_y (\frac{q_x q_y}{\rho}) - \epsilon_1 \partial_{xx}(q_x) = 0 \\
&\partial_t q_y + \partial_x (\frac{q_x q_y}{\rho}) + \partial_y (\frac{q_y^2}{\rho} + \frac{g\rho^2}{2}) - \epsilon_2 \partial_{yy}(q_y) = 0 \\
&\rho|_{t=0} = \rho_0, \quad q_x|_{t=0} = q_{x_0}, \quad q_y|_{t=0} = q_{y_0} \\
&\rho = \rho^b, \quad q_x = q_x^b, \quad q_y = q_y^b \quad on \quad \partial\Omega \times (0,T).
\end{aligned} \tag{6.1}
$$

where $\rho_0, q_{x_0}, q_{y_0}, \rho^b, q_x^b, q_y^b$ are given. Different types of boundary condition can be chosen. We have chosen to use Dirichlet boundary condition in our analysis.

### 6.1.1 Assumptions

For the notations please refer to chapter 3.

- (i) We assume that the numerical flux $H(u, v, n)$ corresponding to the advective flux $f_s$ is Lipschitz continuous, i.e.

$$|H(u, v, n) - H(u^*, v^*, n)| \leq L_f(|u - u^*| + |v - v^*|) \quad \forall u, v, u^*, v^* \in \mathbb{R}. \tag{6.2}$$

  $H(u, v, n)$ is consistent, i.e.

$$H(u, u, n) = \sum_{s=1}^{d} f_s(u) n_s. \tag{6.3}$$

  $H(u, v, n)$ is conservative, i.e.

$$H(u, v, n) = -H(v, u, -n). \tag{6.4}$$

  where $L_f$ is the Lipschitz constant. From (6.2) and (6.3) we can conclude that the functions $f_s$ are also Lipschitz continuous.

- (ii) The diffusion coefficient is assumed to be bounded, i.e.

$$0 < \epsilon_\circ \leq \epsilon \leq \epsilon^\circ.$$

  where $\epsilon$ can be $\epsilon_1$ or $\epsilon_2$.

- (iii) (Compatibility condition) It is assumed that, for an element, the restriction of the finite element space of the solution $(\rho_h, q_{x_h}, q_{y_h})$ to the

boundary of the element is a subset of the corresponding finite element space of the trace elements $(\lambda_{\rho_h}, \lambda_{q_{x_h}}, \lambda_{q_{y_h}})$, i.e.

$$\rho_h|_{\partial K} \subset \lambda_{\rho_h}, \quad q_{x_h}|_{\partial K} \subset \lambda_{q_{x_h}}, \quad q_{y_h}|_{\partial K} \subset \lambda_{q_{y_h}}.$$

- (iv) It is assumed that the space of $\rho$ is a subspace of the space of the components of $\sigma_x$ and $\sigma_y$.

## 6.2 Space-time projection

The time discretization is based on a partition of $0 = t_0 < t_1 < ... < t_M = T$ of the time interval $I = [0, T]$ into subintervals $I_m = (t_{m-1}, t_m)$ of length $\tau_m = t_m - t_{m-1}$. Each space-time slab, $S_m = \Omega \times I_m$ is divided into space-time prisms $K_i \times I_m$, where $\mathfrak{T}_{h,m} = \{K_i\}, i \in I$ is a triangulation of $\Omega$ with mesh discretization parameter $h_m$ and $I$ is an index set. The space mesh may change from one time interval to another (e.g. due to adaptivity). Thus in general there are two space-meshes associated to each time level $t_m$, namely one from below and one from above. In case that these two meshes are not aligned, hanging nodes are inevitable. An interior face "$e$" is any planar set of positive $(d-1)$-dimensional measure of the form $e = \partial K^+ \cap \partial K^-$ for some two elements $K^+, K^- \in \mathfrak{T}_{h,m}$. Similarly, we say that "$e$" is a boundary face if $e = \partial K^+ \cap \partial\Omega$ and the $(d-1)$-dimensional Lebesgue measure of "$e$" is not zero. We define $\Gamma_{h,m} = \{e\}_i, i \in J$ as the set of all faces of the mesh where $J$ is an index set. By $\Gamma^{int}_{h,m} = \{e\}_i, i \in J^{int}$ and $\Gamma^{b}_{h,m} = \{e\}_i, i \in J^{b}$ we mean the set of all interior and boundary faces, respectively. Therefore

$J = J^{int} \cup J^b$. The set of boundary faces is also defined to be the union of Dirichlet ($\Gamma_{h,m}^{b,D} = \{e\}_i, i \in J_D^b$) and Neumann ($\Gamma_{h,m}^{b,N} = \{e\}_i, i \in J_N^b$) parts of the boundary. We define $\Xi_T = \bigcup_m \Gamma_{h,m} \times I_m$ and $Q_T = \bigcup_m \mathcal{T}_{h,m} \times I_m$. As the solution is discontinuous between time slabs we also define

$$\phi_m^{\pm} = \lim_{\epsilon>0,\epsilon\to 0} \phi(t_m \pm \epsilon), \quad [\![\phi]\!]_m = \phi_m^+ - \phi_m^-.$$

for an arbitrary function $\phi$. Over a triangulation $\mathcal{T}_{h,m}$, the following finite dimensional broken Sobolev and Bochner spaces are defined:

$$S_{h,\tau}^{p,r} = \{\phi(x,t) : \phi(x,t) \in L^2(\Omega \times [0,T]) : \phi|_{I_m \times K} = \mathcal{P}^r(I_m; \mathcal{P}^p(K))\}$$
$$S_{h,m}^{p} = \{\phi(x) : \phi(x) \in L^2(\Omega) : \phi|_K = \mathcal{P}^p(K) \ \forall K \in \mathcal{T}_{h,m}\}$$
$$M_{h,\tau}^{p,r} = \{\phi(x,t) : \phi(x,t) \in L^2(\Xi_T) : \phi|_{I_m \times e} = \mathcal{P}^r(I_m; \mathcal{P}^p(e)) \ \}$$
$$M_{h,\tau}^{p,r}(g_D) = \{\phi \in M_{h,\tau}^{p,q} : \phi|_{\Gamma_{h,m}^b} = P g_D\},$$

where by $\mathcal{P}^r(I_m; \mathcal{P}^p(K))$ we mean the space of polynomials of order $r$ in time with values in the polynomial space $\mathcal{P}^p(K)$ and $P$ is an $L^2$ projection. Note that as we are dealing with different mesh partition at a discrete time level $t_m$, we can not simply upwind the solution from previous time slab to the next one as this might introduce discontinuity inside the space-time element of the current slab and thus cause the solution not to be in space $S_{h,\tau}^{p,r}$. Therefore the solution at the end of previous time slab must be $L^2$ projected to the space of the solution at the current time slab.

For the error estimate we need to define a space-time projection operator as follows:

**Definition 6.2.1.** The space-time projection operator $\pi$ acting on elements of $v \in H^1(0,T;L^2)$ is defined as follows [34]:

$$\begin{aligned} &\pi v \in S_{h,\tau}^{p,r} \\ &\pi v(t_m^-) = \Pi_m v(t_m^-) \\ &\int_{I_m} (\pi v - v, \phi) dt = 0, \quad \forall \phi \in S_{h,\tau}^{p,r-1}, \quad m = 1, ...M \end{aligned} \tag{6.5}$$

where $\Pi_m$ is an $L^2$ projection on $S_{h,m}^p$, i.e.

$$(\Pi_m u - u, \phi) = 0, \quad \forall \phi \in S_{h,m}^p. \tag{6.6}$$

We are essentially using Gauss-Radau projection in time to be compatible with our chosen space as we are using upwinding in time and thus have continuity in the time direction. The trace values are not continuous across the time slabs, and hence we are going to define the usual $L^2$ projection for them, i.e.

**Definition 6.2.2.** We define a space-time projection $\pi'$ acting on an element $v \in L^2(0,T;L^2)$ such that

$$\begin{aligned} &\pi' v \in M_{h,\tau}^{p,r} \\ &\int_{I_m} (\pi' v - v, \phi) dt = 0, \quad \forall \phi \in M_{h,\tau}^{p,r}, \quad m = 1, ...M. \end{aligned}$$

We also introduce the following notations:

$$
\begin{aligned}
(u,v) &:= \sum_{K_i \in \mathcal{T}_{h,m}} \int_{K_i} uv dx \\
\langle u,v \rangle &:= \sum_{K_i \in \mathcal{T}_{h,m}} \int_{\partial K_i} uv ds \\
\|.\| &:= \sum_{K_i \in \mathcal{T}_{h,m}} \|.\|_{L^2(K_i)} \\
\|.\|_{H^1} &:= \sum_{K_i \in \mathcal{T}_{h,m}} \|.\|_{H^1(K_i)} \\
\|.\|_{L^2(\Gamma_{h,m})} &:= \sum_{K_i \in \mathcal{T}_{h,m}} \|.\|_{L^2(\partial K_i)}.
\end{aligned}
$$

The meaning of other similar notations may be inferred from these. We use upwinding in time and hybridization in space.

## 6.3 Abstract error estimate

Multiplying the equations (6.1) by test functions in $S^{p,r}_{h,\tau}$ and $M^{p,r}_{h,\tau}$, integrating over an element $K \times I_m$ and summing over all elements, we obtain: Find $\rho, q_x, q_y, \sigma_x, \sigma_y, \lambda_\rho, \lambda_{q_x}, \lambda_{q_y} \in S^{p,r}_{h,\tau} \times S^{p,r}_{h,\tau} \times S^{p,r}_{h,\tau} \times (S^{p,r}_{h,\tau})^2 \times (S^{p,r}_{h,\tau})^2 \times M^{p,r}_{h,\tau}(\rho^b) \times S^{p,r}_{h,\tau}(q^b_x) \times S^{p,r}_{h,\tau}(q^b_y)$ such that

$$
\begin{aligned}
&\int_{I_m} (\frac{\partial \rho}{\partial t}, w) dt + ((\rho^+_{m-1} - \rho^-_{m-1}), w^+_{m-1}) + \int_{I_m} \Big( \langle \lambda_{q_x}, n_x w \rangle - (q_x, \frac{\partial w}{\partial x}) + \\
&\langle \lambda_{q_y}, n_y w \rangle - (q_y, \frac{\partial w}{\partial y}) + \alpha \langle (\rho - \lambda_\rho), w \rangle \Big) dt = 0, \quad \forall w \in S^{p,r}_{h,\tau}, \qquad (6.7)
\end{aligned}
$$

$$\int_{I_m}(\frac{\partial q_x}{\partial t},z)dt+((q^+_{x_{m-1}}-q^-_{x_{m-1}}),z^+_{m-1})+\int_{I_m}\Big(\langle\frac{\lambda^2_{q_x}}{\lambda_\rho}+g\frac{\lambda^2_\rho}{2},zn_x\rangle$$
$$-(\frac{{q_x}^2}{\rho}+\frac{g\rho^2}{2},\partial_x z)+\langle\frac{\lambda_{q_y}\lambda_{q_x}}{\lambda_\rho},zn_y\rangle-(\frac{q_xq_y}{\rho},\partial_y z)+\beta\langle(q_x-\lambda_{q_x}),z\rangle$$
$$+\langle\sigma_x\cdot n,z\rangle-(\sigma_x,\nabla z)\Big)dt=0,\quad\forall z\in S^{p,r}_{h,\tau},\tag{6.8}$$

$$\int_{I_m}(\frac{\partial q_y}{\partial t},l)dt+((q^+_{y_{m-1}}-q^-_{y_{m-1}}),l^+_{m-1})+\int_{I_m}\Big(\langle\frac{\lambda^2_{q_y}}{\lambda_\rho}+g\frac{\lambda^2_\rho}{2},ln_y\rangle$$
$$-(\frac{{q_y}^2}{\rho}+\frac{g\rho^2}{2},\partial_y l)+\langle\frac{\lambda_{q_y}\lambda_{q_x}}{\lambda_\rho},ln_x\rangle-(\frac{q_xq_y}{\rho},\partial_x l)+\beta'\langle(q_y-\lambda_{q_y}),l\rangle$$
$$+\langle\sigma_y\cdot n,l\rangle-(\sigma_y,\nabla l)\Big)dt=0,\quad\forall l\in S^{p,r}_{h,\tau},\tag{6.9}$$

$$\int_{I_m}\Big(\frac{1}{\epsilon_1}(\sigma_x,\tau_1)+\langle\lambda_{q_x},\tau_1\cdot n\rangle-(q_x,\nabla\cdot\tau_1)\Big)dt=0,\quad\forall\tau_1\in(S^{p,r}_{h,\tau})^2,\tag{6.10}$$
$$\int_{I_m}\Big(\frac{1}{\epsilon_2}(\sigma_y,\tau_2)+\langle\lambda_{q_y},\tau_2\cdot n\rangle-(q_y,\nabla\cdot\tau_2)\Big)dt=0,\quad\forall\tau_2\in(S^{p,r}_{h,\tau})^2,\tag{6.11}$$

$$\int_{I_m}\Big(\alpha\langle(\rho-\lambda_\rho),\mu_1\rangle\Big)dt=0,\qquad\forall\mu_1\in M^{p,r}_{h,\tau}(0),\tag{6.12}$$
$$\int_{I_m}\Big(\langle\sigma_x\cdot n,\mu_2\rangle+\beta\langle(q_x-\lambda_{q_x}),\mu_2\rangle\Big)dt=0,\qquad\forall\mu_2\in M^{p,r}_{h,\tau}(0),\tag{6.13}$$
$$\int_{I_m}\Big(\langle\sigma_y\cdot n,\mu_3\rangle+\beta'\langle(q_y-\lambda_{q_y}),\mu_3\rangle\Big)dt=0,\qquad\forall\mu_3\in M^{p,r}_{h,\tau}(0),\tag{6.14}$$

where $\sigma_x=-\epsilon_1\nabla q_x$, $\sigma_y=-\epsilon_2\nabla q_y$, $\alpha,\beta,\beta'$ are the local stabilization coefficients.

Decomposing the error corresponding to each variable, we obtain:

$$e_\rho = \rho - \rho_{h\tau} = (\rho - \pi\rho) - (\rho_{h\tau} - \pi\rho) = \eta_\rho - \xi_\rho,$$
$$e_{\lambda_\rho} = \lambda_\rho - \lambda_{\rho_{h\tau}} = (\lambda_\rho - \pi^{'}\lambda_\rho) - (\lambda_{\rho_{h\tau}} - \pi^{'}\lambda_\rho) = \eta_{\lambda_\rho} - \xi_{\lambda_\rho}$$

and the same for the rest of variables. The subindex $h\tau$ is for the finite element solution. As we have $L^2$ projected the boundary condition onto the space of trace elements, we have $\xi_{\lambda_\rho} = \xi_{\lambda_{q_x}} = \xi_{\lambda_{q_y}} = 0$ on the boundary.

We now obtain the error equations. For brevity, the error equation for one of the flux, namely $q_x$ will be written. The same procedure can be repeated for $q_y$. We start with the error equations for the conservativity conditions (6.12) and (6.13):

$$\int_{I_m} \alpha\langle(\xi_\rho - \xi_{\lambda_\rho}), \mu_1\rangle dt = \int_{I_m} \alpha\langle(\eta_\rho - \eta_{\lambda_\rho}), \mu_1\rangle dt \tag{6.15}$$

$$\int_{I_m} \big(\langle\xi_{\sigma_x}\cdot n, \mu_2\rangle + \beta\langle(\xi_{q_x} - \xi_{\lambda_{qx}}), \mu_2\rangle\big)dt = \int_{I_m} \big(\langle\eta_{\sigma_x}\cdot n, \mu_2\rangle + \beta\langle(\eta_{q_x} - \eta_{\lambda_{qx}}), \mu_2\rangle\big)dt \tag{6.16}$$

Before writing the error equation for (6.7), we are first going to simplify it. Based on assumption (iv) we can choose $(w, 0)$ as a test function in (6.10) and $(0, w)$ as a test function in (6.11). Therefore (6.7) can be written as

$$\int_{I_m} (\frac{\partial\rho}{\partial t}, w)dt + ((\rho^+_{m-1} - \rho^-_{m-1}), w^+_{m-1}) + \int_{I_m} \Big( -\frac{1}{\epsilon_1}(\sigma_{x1}, w) - \frac{1}{\epsilon_2}(\sigma_{y2}, w) + \alpha\langle(\rho - \lambda_\rho), w\rangle\Big)dt = 0 \quad \forall w \in S^{p,r}_{h,\tau} \tag{6.17}$$

Using $\xi_\rho$ in (6.17), $-\xi_{\lambda_\rho}$ in (6.15) as test functions and then summing up the results, the error equation for (6.17) will be

$$\int_{I_m}(\frac{\partial \xi_\rho}{\partial t},\xi_\rho)dt+((\xi^+_{\rho,m-1}-\xi^-_{\rho,m-1}),\xi^+_{\rho,m-1})+\int_{I_m}\Big(-\frac{1}{\epsilon_1}(\xi_{\sigma_{x_1}},\xi_\rho)-\frac{1}{\epsilon_2}(\xi_{\sigma_{y_2}},\xi_\rho)$$
$$+\alpha\langle(\xi_\rho-\xi_{\lambda_\rho}),(\xi_\rho-\xi_{\lambda_\rho})\rangle\Big)dt$$
$$=\int_{I_m}(\frac{\partial \eta_\rho}{\partial t},\xi_\rho)dt+((\eta^+_{\rho,m-1}-\eta^-_{\rho,m-1}),\xi^+_{\rho,m-1})+\int_{I_m}\Big(-\frac{1}{\epsilon_1}(\eta_{\sigma_{x_1}},\xi_\rho)-\frac{1}{\epsilon_2}(\eta_{\sigma_{y_2}},\xi_\rho)$$
$$+\alpha\langle(\eta_\rho-\eta_{\lambda_\rho}),(\xi_\rho-\xi_{\lambda_\rho})\rangle\Big)dt. \tag{6.18}$$

The first two terms on the left hand side can be simplified in two different ways, either as

$$\int_{I_m}(\frac{\partial \xi_\rho}{\partial t},\xi_\rho)dt+((\xi^+_{\rho,m-1}-\xi^-_{\rho,m-1}),\xi^+_{\rho,m-1})=\frac{1}{2}(\|\xi^-_{\rho,m}\|^2-\|\xi^+_{\rho,m-1}\|^2)$$
$$+\|\xi^+_{\rho,m-1}\|^2-(\xi^-_{\rho,m-1},\xi^+_{\rho,m-1})=\frac{1}{2}(\|\xi^-_{\rho,m}\|^2+\|\xi^+_{\rho,m-1}\|^2)-(\xi^-_{\rho,m-1},\xi^+_{\rho,m-1}) \tag{6.19}$$

or

$$2\times\Big(\int_{I_m}(\frac{\partial \xi_\rho}{\partial t},\xi_\rho)dt+((\xi^+_{\rho,m-1}-\xi^-_{\rho,m-1}),\xi^+_{\rho,m-1})\Big)=\|\xi^-_{\rho,m}\|^2-\|\xi^+_{\rho,m-1}\|^2$$
$$+2((\xi^+_{\rho,m-1}-\xi^-_{\rho,m-1}),\xi^+_{\rho,m-1}))=\|\xi^-_{\rho,m}\|^2-\|\xi^+_{\rho,m-1}\|^2+\|\xi^+_{\rho,m-1}\|^2-(\xi^-_{\rho,m-1},\xi^+_{\rho,m-1})$$
$$+(\xi^+_{\rho,m-1}-\xi^-_{\rho,m-1},\xi^+_{\rho,m-1}-\xi^-_{\rho,m-1})+(\xi^+_{\rho,m-1}-\xi^-_{\rho,m-1},\xi^-_{\rho,m-1}),$$

which can be simplified as

$$\int_{I_m}(\frac{\partial \xi_\rho}{\partial t},\xi_\rho)dt+((\xi^+_{\rho,m-1}-\xi^-_{\rho,m-1}),\xi^+_{\rho,m-1})=$$
$$\frac{1}{2}(\|\xi^-_{\rho,m}\|^2+\|\xi^+_{\rho,m-1}-\xi^-_{\rho,m-1}\|^2-\|\xi^-_{\rho,m-1}\|^2)=\frac{1}{2}(\|\xi^-_{\rho,m}\|^2-\|\xi^-_{\rho,m-1}\|^2)+\frac{1}{2}\|[\![\xi]\!]_{\rho,m-1}\|^2. \tag{6.20}$$

Both forms are going to be used. For now we use the form (6.19).

The first term on the right hand side of (6.18) can be simplified as

$$\int_{I_m} (\frac{\partial \eta_\rho}{\partial t}, \xi_\rho) dt = -(\eta^+_{\rho,m-1}, \xi^+_{\rho,m-1}) + (\eta^-_{\rho,m}, \xi^-_{\rho,m}) - \int_{I_m} (\eta_\rho, \frac{\partial \xi_\rho}{\partial t}) dt$$

where the last two terms are zero based on definition of the projection (6.5). Therefore

$$\int_{I_m} (\frac{\partial \eta_\rho}{\partial t}, \xi_\rho) dt + ((\eta^+_{\rho,m-1} - \eta^-_{\rho,m-1}), \xi^+_{\rho,m-1}) = -(\eta^-_{\rho,m-1}, \xi^+_{\rho,m-1}).$$

This term can be bounded either as

$$(\eta^-_{\rho,m-1}, \xi^+_{\rho,m-1}) \leq \frac{1}{4\delta} \|\eta^-_{\rho,m-1}\|^2 + \delta \|\xi^+_{\rho,m-1}\|^2 \tag{6.21}$$

or

$$(\eta^-_{\rho,m-1}, \xi^+_{\rho,m-1}) = (\eta^-_{\rho,m-1}, \xi^+_{\rho,m-1} - \xi^-_{\rho,m-1}) \leq \frac{1}{4\delta} \|\eta^-_{\rho,m-1}\|^2 + \delta \|[\![\xi]\!]_{\rho,m-1}\|^2 \tag{6.22}$$

for a positive generic constant $\delta$.

Before bounding the other terms in (6.18), we define the trace and inverse inequalities as follows:

**Lemma 6.3.1.** *(Trace inequality) For a function $v \in H^1(K)$, there exists a constant $C_M$ such that*

$$\|v\|^2_{L^2(\partial K)} \leq C_M \big( \|v\|^2_{L^2(K)} |v|_{H^1(K)} + h_K^{-1} \|v\|_{L^2(K)} \big) \tag{6.23}$$

**Lemma 6.3.2.** *(Inverse inequality) There exists a constant $C_I$ such that*

$$|v|_{H^1(K)} \leq C_I h_K^{-1} \|v\|_{L^2(K)} \quad v \in \mathcal{P}^p(K). \tag{6.24}$$

For the proof, the reader can refer to any standard finite element text (see, e.g. [8]).

Combining (6.23) and (6.24), we have the following:

$$\|v\|^2_{L^2(\partial K)} \leq c_{tr} h_K^{-1} \|v\|_{L^2(K)} \quad v \in \mathcal{P}^p(K) \tag{6.25}$$

Using (6.25), the other terms in (6.18) can be bounded as follows:

$$\alpha\langle(\eta_\rho - \eta_{\lambda_\rho})n_x, (\xi_\rho - \xi_{\lambda_\rho})\rangle \leq \frac{\alpha}{2}\|\eta_\rho - \eta_{\lambda_\rho}\|^2_{L^2(\Gamma_{h,n})} + \frac{\alpha}{2}\|\xi_\rho - \xi_{\lambda_\rho}\|^2_{L^2(\Gamma_{h,n})},$$

where the second term can be absorbed in the left hand side of equation (6.18). Also

$$\begin{aligned}\frac{1}{\epsilon_1}(\xi_{\sigma_{x_1}}, \xi_\rho) + \frac{1}{\epsilon_2}(\xi_{\sigma_{y_2}}, \xi_\rho) &\leq \frac{\delta_2}{\epsilon_1}\|\xi_{\sigma_x}\|^2 + \frac{\delta_2}{\epsilon_2}\|\xi_{\sigma_y}\|^2 + \frac{1}{4\delta_2}(\frac{1}{\epsilon_1} + \frac{1}{\epsilon_2})\|\xi_\rho\|^2 \\ &\leq \frac{\delta_2}{\epsilon_1}\|\xi_{\sigma_x}\|^2 + \frac{\delta_2}{\epsilon_2}\|\xi_{\sigma_y}\|^2 + c'\|\xi_\rho\|^2\end{aligned}$$

where

$$c' = \frac{1}{4\delta_2}(\frac{1}{\epsilon_{1\circ}} + \frac{1}{\epsilon_{2\circ}}) > \frac{1}{4\delta_2}(\frac{1}{\epsilon_1} + \frac{1}{\epsilon_2}).$$

We are not going to keep track of the dependence of the constants like $c'$ in above on $\epsilon_{1_0}$ and $\epsilon_{2_0}$. Bounding the other terms similarly and using (6.19) and (6.21) we obtain

$$\begin{aligned}&\frac{1}{2}(\|\xi^-_{\rho,m}\|^2 + \|\xi^+_{\rho,m-1}\|^2) + \int_{I_m}\Big(\frac{\alpha}{2}\langle(\xi_\rho - \xi_{\lambda_\rho}), (\xi_\rho - \xi_{\lambda_\rho})\rangle\Big)dt \\ &\leq \frac{1}{4\delta_1}\|\eta^-_{\rho,m-1}\|^2 + \delta_1\|\xi^+_{\rho,m-1}\|^2 + \int_{I_m}\Big(c_1\|\xi_\rho\|^2 + \frac{\alpha}{2}\|\eta_\rho - \eta_{\lambda_\rho}\|^2_{L^2(\Gamma_{h,n})} \\ &+ \frac{\delta_2}{\epsilon_1}\|\xi_{\sigma_x}\|^2 + \frac{\delta_2}{\epsilon_2}\|\xi_{\sigma_y}\|^2 + \frac{c_2}{\epsilon_1}\|\eta_{\sigma_x}\|^2 + \frac{c_3}{\epsilon_2}\|\eta_{\sigma_y}\|^2\Big)dt + \frac{1}{4\delta_3}\|\xi^-_{\rho,m-1}\|^2 + \delta_3\|\xi^+_{\rho,m-1}\|^2.\end{aligned} \tag{6.26}$$

Before writing the error equation for the fluxes we first obtain the error equation for (6.10). Using $\xi_{\sigma_x}$ as a test function and doing integration by parts we obtain

$$\int_{I_m} \Big(\frac{1}{\epsilon_1}(\xi_{\sigma_x}, \xi_{\sigma_x}) + \langle(\xi_{\lambda_{q_x}} - \xi_{q_x}), \xi_{\sigma_x} \cdot n\rangle + (\nabla\xi_{q_x}, \xi_{\sigma_x})\Big)dt = \tag{6.27}$$
$$\int_{I_m} \Big(\frac{1}{\epsilon_1}(\eta_{\sigma_x}, \xi_{\sigma_x}) + \langle(\eta_{\lambda_{q_x}} - \eta_{q_x}), \xi_{\sigma_x} \cdot n\rangle + (\nabla\eta_{q_x}, \xi_{\sigma_x})\Big)dt$$

By simplifying the right hand side and using the trace inequality (6.25) we have

$$\int_{I_m} \Big(\frac{1}{\epsilon_1}(\xi_{\sigma_x}, \xi_{\sigma_x}) + < (\xi_{\lambda_{q_x}} - \xi_{q_x}), \xi_{\sigma_x} \cdot n > + (\nabla\xi_{q_x}, \xi_{\sigma_x})\Big)dt \leq$$
$$\frac{c_1}{\epsilon_1}\|\eta_\sigma\|^2 + \frac{c_2\epsilon_1}{h}\|\eta_{\lambda_{q_x}} - \eta_{q_x}\|^2_{L^2(\Gamma_{h,n})} + \frac{\delta}{\epsilon_1}\|\xi_{\sigma_x}\|^2 + \epsilon_1^\circ c_3\|\nabla\eta_{q_x}\|^2. \tag{6.28}$$

where $\delta$ can be chosen as small as we wish.

The error equation for the flux in $x$ direction is now obtained by using $\xi_{q_x}$ and $-\xi_{\lambda_{q_x}}$ in (6.8) and (6.16) respectively and then summing up the results, i.e.

$$\int_{I_m} (\frac{\partial\xi_{q_x}}{\partial t}, \xi_{q_x})dt + ((\xi^+_{q_{x_{m-1}}} - \xi^-_{q_{x_{m-1}}}), \xi^+_{q_{x_{m-1}}})$$
$$+ \int_{I_m} \Big(\langle\xi_{\sigma_x} \cdot n, \xi_{q_x} - \xi_{\lambda_{q_x}}\rangle - (\xi_{\sigma_x}, \nabla\xi_{q_x}) + \beta\langle(\xi_{q_x} - \xi_{\lambda_{q_x}}), (\xi_{q_x} - \xi_{\lambda_{q_x}})\rangle\Big)dt =$$
$$\int_{I_m} (\frac{\partial\eta_{q_x}}{\partial t}, \xi_{q_x})dt + ((\eta^+_{q_{x_{m-1}}} - \eta^-_{q_{x_{m-1}}}), \eta^+_{q_{x_{m-1}}}) + \int_{I_m} \Big(\langle\eta_{\sigma_x} \cdot n, \xi_{q_x} - \xi_{\lambda_{q_x}}\rangle$$
$$- (\eta_{\sigma_x}, \nabla\xi_{q_x}) - \langle\frac{\lambda^2_{q_x}}{\lambda_\rho} + g\frac{\lambda^2_\rho}{2}, \xi_{q_x} n_x\rangle + \langle\frac{\lambda^2_{q_{x_{h\tau}}}}{\lambda_{\rho_{h\tau}}} + g\frac{\lambda^2_{\rho_{h\tau}}}{2}, \xi_{q_x} n_x\rangle + (\frac{{q_x}^2}{\rho} + \frac{g\rho^2}{2}, \partial_x\xi_{q_x})$$
$$- (\frac{{q_{x_{h\tau}}}^2}{\rho_{h\tau}} + \frac{g\rho^2_{h\tau}}{2}, \partial_x\xi_{q_x}) - \langle\frac{\lambda_{q_y}\lambda_{q_x}}{\lambda_\rho}, \xi_{\lambda_{q_x}} n_y\rangle + \langle\frac{\lambda_{q_{y_{h\tau}}}\lambda_{q_{x_{h\tau}}}}{\lambda_{\rho_{h\tau}}}, \xi_{q_x} n_y\rangle + (\frac{q_x q_y}{\rho}, \partial_y\xi_{q_x})$$
$$- (\frac{q_{x_{h\tau}} q_{y_{h\tau}}}{\rho_{h\tau}}, \partial_y\xi_{q_x}) + \beta\langle(\eta_{q_x} - \eta_{\lambda_{q_x}}), (\xi_{q_x} - \xi_{\lambda_{q_x}})\rangle\Big)dt$$

Now using the Lipschitz continuity of the advective flux (6.2), the error equation can be simplified as

$$\int_{I_m}(\frac{\partial \xi_{q_x}}{\partial t},\xi_{q_x})dt+((\xi^+_{q_x,m-1}-\xi^-_{q_x,m-1}),\xi^+_{q_x,m-1})+\int_{I_m}\Big(-(\xi_{\sigma_x},\nabla\xi_{q_x})$$
$$+\langle\xi_{\sigma_x}\cdot n,\xi_{q_x}-\xi_{\lambda_{q_x}}\rangle+\beta\langle(\xi_{q_x}-\xi_{\lambda_{q_x}}),\xi_{q_x}-\xi_{\lambda_{q_x}}\rangle\Big)dt\leq$$
$$\int_{I_m}\Big(L_1\langle|\xi_{\lambda_{q_x}}||n_x|,|\xi_{q_x}|\rangle+L_2\langle|\xi_{\lambda_\rho}||n_x|,|\xi_{q_x}|\rangle$$
$$+L_3(|\xi_{q_x}|,|\frac{\partial\xi_{q_x}}{\partial x}|)+L_4(|\xi_\rho|,|\frac{\partial\xi_{q_x}}{\partial x}|)+L_5\langle|\xi_{\lambda_{q_x}}||n_y|,|\xi_{q_x}|\rangle+L_6\langle|\xi_{\lambda_\rho}||n_y|,|\xi_{q_x}|\rangle$$
$$+L_7\langle|\xi_{\lambda_{q_y}}||n_y|,|\xi_{q_x}|\rangle+L_8(|\xi_{q_x}|,|\frac{\partial\xi_{q_x}}{\partial y}|)+L_9(|\xi_\rho|,|\frac{\partial\xi_{q_x}}{\partial y}|)+L_{10}(|\xi_{q_y}|,|\frac{\partial\xi_{q_x}}{\partial y}|)\Big)dt$$
$$+\int_{I_m}(\frac{\partial\eta_{q_x}}{\partial t},\xi_{q_x})dt+((\eta^+_{q_x,m-1}-\eta^-_{q_x,m-1}),\xi^+_{q_x,m-1})+\int_{I_m}\Big(L_1\langle|\eta_{\lambda_{q_x}}||n_x|,|\xi_{q_x}|\rangle$$
$$+L_2\langle|\eta_{\lambda_\rho}||n_x|,|\xi_{q_x}|\rangle+L_3(|\eta_{q_x}|,|\frac{\partial\xi_{q_x}}{\partial x}|)+L_4(|\eta_\rho|,|\frac{\partial\xi_{q_x}}{\partial x}|)+L_5\langle|\eta_{\lambda_{q_x}}||n_y|,|\xi_{q_x}|\rangle$$
$$+L_6\langle|\eta_{\lambda_\rho}||n_y|,|\xi_{q_x}|\rangle+L_7\langle|\eta_{\lambda_{q_y}}||n_y|,|\xi_{q_x}|\rangle+L_8(|\eta_{q_x}|,|\frac{\partial\xi_{q_x}}{\partial y}|)+L_9(|\eta_\rho|,|\frac{\partial\xi_{q_x}}{\partial y}|)$$
$$+L_{10}(|\eta_{q_y}|,|\frac{\partial\xi_{q_x}}{\partial y}|)-(\eta_{\sigma_x},\nabla\xi_{q_x})+\langle\eta_{\sigma_x}\cdot n,\xi_{q_x}-\xi_{\lambda_{q_x}}\rangle$$
$$+\beta\langle(\eta_{q_x}-\eta_{\lambda_{q_x}}),\xi_{q_x}-\xi_{\lambda_{q_x}}\rangle\Big)dt, \tag{6.29}$$

where $L_1, L_2, ...L_{10}$ are the Lipschitz constants and we have used (6.2) obtain these terms, e.g.

$$\langle\frac{\lambda^2_{q_x}}{\lambda_\rho}+g\frac{\lambda^2_\rho}{2},zn_x\rangle-\langle\frac{\lambda^2_{q_{x_{h\tau}}}}{\lambda_{\rho_{h\tau}}}+g\frac{\lambda^2_{\rho_{h\tau}}}{2},zn_x\rangle\leq\langle|\frac{\lambda^2_{q_x}}{\lambda_\rho}+g\frac{\lambda^2_\rho}{2}-(\frac{\lambda^2_{q_{x_{h\tau}}}}{\lambda_{\rho_{h\tau}}}+g\frac{\lambda^2_{\rho_{h\tau}}}{2})|,|z||n_x|\rangle$$
$$\leq L\langle|\lambda_{q_x}-\lambda_{q_{x_{h\tau}}}|,|z||n_x|\rangle+L'\langle|\lambda_\rho-\lambda_{\rho_{h\tau}}|,|z||n_x|\rangle\leq L\langle|\xi_{\lambda_{q_x}}|,|z||n_x|\rangle$$
$$+L\langle|\eta_{\lambda_{q_x}}|,|z||n_x|\rangle+L'\langle|\xi_{\lambda_\rho}|,|z||n_x|\rangle+L'\langle|\eta_{\lambda_\rho}|,|z||n_x|\rangle.$$

The first two terms on the left hand side of (6.29) can be bounded

similar to the error equation for $\rho$, the other terms can be treated as follows:

$$\begin{aligned}\langle|\xi_{\lambda_\rho}||n_x|,|\xi_{q_x}|\rangle \leq \langle|\xi_{\lambda_\rho}|,|\xi_{q_x}|\rangle \quad &= \langle|\xi_{\lambda_\rho}|-|\xi_\rho|,|\xi_{q_x}|\rangle + \langle|\xi_\rho|,|\xi_{q_x}|\rangle \\ &\leq \langle\frac{1}{\sqrt{h}}|\xi_{\lambda_\rho}-\xi_\rho|,\sqrt{h}|\xi_{q_x}|\rangle + \langle\sqrt{\frac{h}{\epsilon_1}}|\xi_\rho|,\sqrt{\frac{\epsilon_1}{h}}|\xi_{q_x}|\rangle \\ &\leq \frac{c_1}{h}\|\xi_{\lambda_\rho}-\xi_\rho\|^2_{L^2(\Gamma_{h,n})} + c_2\|\xi_{q_x}\|^2 + \frac{c_3}{\epsilon_{1\circ}}\|\xi_\rho\|^2 \\ &+ \delta\epsilon_1\|\nabla\xi_{q_x}\|^2\end{aligned}$$

and

$$(|\xi_\rho|,|\frac{\partial\xi_{q_x}}{\partial x}|) \leq \frac{c_1}{\epsilon_{1\circ}}\|\xi_\rho\|^2 + \delta\epsilon_1\|\nabla\xi_{q_x}\|^2$$

and similarly for other terms. Using the above simplifications and adding (6.28) to the result, we end up with

$$\begin{aligned}&\frac{1}{2}(\|\xi^-_{q_x,m}\|^2 + \|\xi^+_{q_x,m-1}\|^2) + \int_{I_m}\Big((\frac{\beta}{2}-\frac{c_0}{h}-c_7)\langle(\xi_{q_x}-\xi_{\lambda_{q_x}}),(\xi_{q_x}-\xi_{\lambda_{q_x}})\rangle \\ &+ (\frac{c_{10}+1/2}{\epsilon_1})\|\xi_{\sigma_x}\|^2\Big)dt \leq \frac{1}{4\delta_1}\|\eta^-_{q_x,m-1}\|^2 + \delta_1\|\xi^+_{q_x,m-1}\|^2 + c_1\int_{I_m}(\|\xi_\rho\|^2 + \|\xi_{q_x}\|^2 \\ &+ \|\xi_{q_y}\|^2)dt + c_2\int_{I_m}(\|\eta_\rho\|^2 + \|\eta_{q_x}\|^2 + \|\eta_{q_y}\|^2 + \|\eta_{\sigma_x}\|^2)dt + \int_{I_m}\Big(\frac{c_3}{h}\|\eta_{\lambda_{q_x}}-\eta_{q_x}\|^2_{L^2(\Gamma_{h,n})} \\ &+ \frac{c_4}{h}\|\eta_{\lambda_\rho}-\eta_\rho\|^2_{L^2(\Gamma_{h,n})} + \delta_2\epsilon_1\|\nabla\xi_{q_x}\|^2 + \delta_2\epsilon_2\|\nabla\xi_{q_y}\|^2 + c_{11}\|\nabla\eta_{q_x}\|^2 + c_{12}\|\nabla\eta_{q_y}\|^2 \\ &+ \frac{\beta}{2}\|\eta_{q_x}-\eta_{\lambda_{q_x}}\|^2_{L^2(\Gamma_{h,n})} + \frac{c_8}{h}\|\eta_{\lambda_{q_y}}-\eta_{q_y}\|^2_{L^2(\Gamma_{h,n})} + \frac{c_9}{h}\|\xi_{\lambda_{q_y}}-\xi_{q_y}\|^2 + \frac{c_{10}}{h}\|\xi_{\lambda_\rho}-\xi_\rho\|^2\Big)dt \\ &+ \frac{1}{4\delta_3}\|\xi^-_{q_x,m-1}\|^2 + \delta_3\|\xi^+_{q_x,m-1}\|^2. \qquad (6.30)\end{aligned}$$

A similar type of estimate can be obtained for $q_y$. As can be seen from (6.30) we need to somehow bound the $\epsilon_1\|\nabla\xi_{q_x}\|^2, \epsilon_2\|\nabla\xi_{q_y}\|^2$ on the right hand side. In order to do this we are going to define a projection similar to the Raviart-Thomas projection but specific to HDG as follows (e.g. see Lemma 3.2 in [15]):

**Lemma 6.3.3.** *Given faces $e_1, ..., e_d$ of a simplex $K \subset \mathbb{R}^d$ and function $\sigma_h \in [L^2(K)]^d$ and $\zeta_{hi} \in L^2(e_i), i = 1, ..., d$, there is a unique function $Z \in [\mathcal{P}^k(K)]^d$ such that*

$$(Z, p)_K = (\sigma_h, p)_K \quad \forall p \in [\mathcal{P}^{k-1}(K)]^d \tag{6.31}$$

$$\langle Z \cdot n_i, \psi \rangle_{e_i} =< \zeta_{hi}, \psi >_{e_i} \quad i = 1, .., d \quad \forall \psi \in \mathcal{P}^k(e_i), \tag{6.32}$$

*where $e_i$ is a face of $K$ and $n_i$ is the unit normal to $e_i, i = 1, ..., d$. We also have the following estimate*

$$\|Z\|_{L^2(K)} \leq C_I \Big( \|\sigma_h\|^2_{L^2(K)} + h_K \sum_{i=1}^{d} \|\zeta_{hi}\|^2_{L^2(e_i)} \Big)^{1/2}, \tag{6.33}$$

*where $C_I$ is a constant depending only on $d, k$ and shape regular constant.*

Note that if the triangulation is composed of simplices, then each simplex has $d + 1$ faces in $d$ dimesion and therefore (6.32) is true on all faces except one. Let us assume that the number of faces of an element $K$ is $n_f$. We choose

$$\begin{aligned} \sigma_h &= \epsilon_1 \nabla \xi_{q_x}, \\ \zeta_{hi} &= \frac{\epsilon_1}{h} (\xi_{\lambda_{q_{xi}}} - \xi_{q_{xi}}) \end{aligned} \tag{6.34}$$

in (6.31) and (6.32) respectively. Therefore (6.33) can be written as

$$\|Z\| \leq C_I (\epsilon_1^2 \|\nabla \xi_{q_x}\|^2 + \frac{\epsilon_1^2}{h} \sum_{i=1}^{d} \|\xi_{\lambda_{q_{xi}}} - \xi_{q_{xi}}\|^2_{L^2(e_i)})^{1/2}. \tag{6.35}$$

We are now going to use $\gamma Z$ ($\gamma$ a constant) as a test function in the error equation of (6.10), after doing integration by parts, we obtain

$$\int_{I_m} \Big(\frac{\gamma}{\epsilon_1}(\xi_{\sigma_x}, Z) + \gamma(\nabla \xi_{q_x}, Z) + \gamma\langle \xi_{\lambda_{q_x}} - \xi_{q_x}, Z \cdot n\rangle\Big) dt =$$
$$\int_{I_m} \Big(\frac{\gamma}{\epsilon_1}(\eta_{\sigma_x}, Z) + \gamma(\nabla \eta_{q_x}, Z) + \gamma\langle \eta_{\lambda_{q_x}} - \eta_{q_x}, Z \cdot n\rangle\Big) dt \tag{6.36}$$

We first focus on the left hand side. The second and the term third can be simplified using (6.34) and choosing $p = \nabla \xi_{q_x}$ in (6.31) and $\psi = (\xi_{\lambda_{q_{xi}}} - \xi_{q_{xi}})$ in (6.32) as follows:

$$\begin{aligned}
&\int_{I_m} \Big(\frac{\gamma}{\epsilon_1}(\xi_{\sigma_x}, Z) + \gamma(\nabla \xi_{q_x}, Z) + \gamma\langle \xi_{\lambda_{q_x}} - \xi_{q_x}, Z \cdot n\rangle\Big) dt \\
&\geq \int_{I_m} \Big( - \frac{1}{2\epsilon_1}\|\xi_{\sigma_x}\|^2 - \frac{\gamma^2}{2\epsilon_1}\|Z\|^2 + \gamma\big(\epsilon_1\|\nabla \xi_{q_x}\|^2 + \frac{\epsilon_1}{h}\sum_{i=1}^{d}\|\xi_{\lambda_{q_x i}} - \xi_{q_{xi}}\|^2_{L^2(e_i)}\big) \\
&+ \sum_{i=d+1}^{n_f} \gamma\langle \xi_{\lambda_{q_{x_i}}} - \xi_{q_{x_i}}, Z \cdot n_{e_i}\rangle_{L^2(e_i)}\Big) dt \\
&\geq \int_{I_m} \Big( - \frac{1}{2\epsilon_1}\|\xi_{\sigma_x}\|^2 + (\gamma - C_I\gamma^2)\big(\epsilon_1\|\nabla \xi_{q_x}\|^2 + \frac{\epsilon_1}{h}\sum_{i=1}^{d}\|\xi_{\lambda_{q_x i}} - \xi_{q_{xi}}\|^2_{L^2(e_i)}\big) \\
&- \frac{c\epsilon_1}{h}\|\xi_{\lambda_{q_x}} - \xi_{q_x}\|^2_{L^2(\partial K)}\Big) dt
\end{aligned} \tag{6.37}$$

where we have used (6.35) and

$$\sum_{i=d+1}^{n_f} \gamma\langle \xi_{\lambda_{q_x}} - \xi_{q_x}, Z \cdot n\rangle_e \geq -\frac{c\epsilon_1}{h}\|\xi_{\lambda_{q_x}} - \xi_{q_x}\|^2_{L^2(\partial K)} - \frac{\gamma^2}{2\epsilon_1}\|Z\|^2.$$

Choosing $\gamma = \frac{1}{2C_I}$, (6.37) can be simplified as

$$\begin{aligned}
&\int_{I_m} \Big(\frac{\gamma}{\epsilon_1}(\xi_{\sigma_x}, Z) + \gamma(\nabla \xi_{q_x}, Z) + \gamma\langle \xi_{\lambda_{q_x}} - \xi_{q_x}, Z \cdot n\rangle\Big) dt \\
&\geq \int_{I_m} \Big( - \frac{1}{2\epsilon_1}\|\xi_{\sigma_x}\|^2 + c_1\epsilon_1\|\nabla \xi_{q_x}\|^2 - \frac{c_2\epsilon_1}{h}\|\xi_{\lambda_{q_x}} - q_x\|^2_{L^2(\Gamma_{h,n})}\Big) dt.
\end{aligned} \tag{6.38}$$

Simplifying the right hand side of (6.36), we obtain

$$\int_{I_m} c'\epsilon_1\|\nabla\xi_{q_x}\|^2dt \leq \int_{I_m}\frac{1}{2\epsilon_1}\|\xi_{\sigma_x}\|^2dt + c_2\int_{I_m}\big(\frac{1}{\epsilon_{1\circ}}\|\eta_{\sigma_x}\|^2$$
$$+\epsilon_1^{\circ}\|\nabla\eta_{q_x}\|^2 + \frac{\epsilon_1^{\circ}}{h}\|\eta_{\lambda_{q_x}}-\eta_{q_x}\|^2 + \frac{c\epsilon_1^{\circ}}{h}\|\xi_{\lambda_{q_x}}-\xi_{q_x}\|^2\big)dt. \tag{6.39}$$

This is a crucial bound that we need for our further analysis. Adding (6.39) to (6.30) we obtain

$$\frac{1}{2}(\|\xi^-_{q_{x,m}}\|^2+\|\xi^+_{q_{x,m-1}}\|^2)+\int_{I_m}\Big((\frac{\beta}{2}-\frac{c_0}{h}-c_7)\langle(\xi_{q_x}-\xi_{\lambda_{q_x}}),(\xi_{q_x}-\xi_{\lambda_{q_x}})\rangle$$
$$(\frac{c_{10}}{\epsilon_1})\|\xi_{\sigma_x}\|^2+c_5\epsilon_1\|\nabla\xi_{q_x}\|^2\Big)dt \leq \frac{1}{4\delta_1}\|\eta^-_{q_{x,m-1}}\|^2+\delta_1\|\xi^+_{q_{x,m-1}}\|^2$$
$$+c_1\int_{I_m}(\|\xi_\rho\|^2+\|\xi_{q_x}\|^2+\|\xi_{q_y}\|^2)dt + c_2\int_{I_m}(\|\eta_\rho\|^2+\|\eta_{q_x}\|^2+\|\eta_{q_y}\|^2+\|\eta_{\sigma_x}\|^2)dt$$
$$+\int_{I_m}\Big(\frac{c_3}{h}\|\eta_{\lambda_{q_x}}-\eta_{q_x}\|^2_{L^2(\Gamma_{h,n})}+\frac{c_4}{h}\|\eta_{\lambda_\rho}-\eta_\rho\|^2_{L^2(\Gamma_{h,n})}+c_{11}\|\nabla\eta_{q_x}\|^2+\frac{\beta}{2}\|\eta_{q_x}-\eta_{\lambda_{q_x}}\|^2_{L^2(\Gamma_{h,n})}$$
$$+\frac{c_8}{h}\|\eta_{\lambda_{q_y}}-\eta_{q_y}\|^2_{L^2(\Gamma_{h,n})}+\frac{c_{10}}{h}\|\xi_{\lambda_\rho}-\xi_\rho\|^2+\frac{c_{11}}{h}\|\xi_{\lambda_{q_y}}-\xi_{q_y}\|^2_{L^2(\Gamma_{h,n})}+\delta_2\epsilon_2\|\nabla\xi_{q_y}\|^2\Big)dt$$
$$+\frac{1}{4\delta_1}\|\xi^-_{q_{x,m-1}}\|^2+\delta_1\|\xi^+_{q_{x,m-1}}\|^2. \tag{6.40}$$

Adding up (6.40) and (6.26) with a similar error inequality for $q_y$ we

obtain

$$\frac{1}{2}(\|\xi^-_{\rho,m}\|^2+\|\xi^+_{\rho,m-1}\|^2)+\int_{I_m}\Big((\frac{\alpha}{2}-\frac{c_0}{h})\langle(\xi_\rho-\xi_{\lambda_\rho}),(\xi_\rho-\xi_{\lambda_\rho})\rangle\Big)dt$$
$$+\frac{1}{2}(\|\xi^-_{q_x,m}\|^2+\|\xi^+_{q_x,m-1}\|^2)+\int_{I_m}\Big((\frac{\beta}{2}-\frac{c_1}{h}-c_2)\langle(\xi_{q_x}-\xi_{\lambda_{q_x}}),(\xi_{q_x}-\xi_{\lambda_{q_x}})\rangle\Big)dt$$
$$+\frac{1}{2}(\|\xi^-_{q_y,m}\|^2+\|\xi^+_{q_y,m-1}\|^2)+\int_{I_m}\Big((\frac{\beta'}{2}-\frac{c_4}{h}-c_5)\langle(\xi_{q_y}-\xi_{\lambda_{q_y}}),(\xi_{q_y}-\xi_{\lambda_{q_y}})\rangle\Big)dt$$
$$+\int_{I_m}\Big(c_7\epsilon_1\|\nabla\xi_{q_x}\|^2+c_8\epsilon_2\|\nabla\xi_{q_y}\|^2+(\frac{c_3}{\epsilon_1})\|\xi_{\sigma_x}\|^2+(\frac{c_6}{\epsilon_2})\|\xi_{\sigma_y}\|^2\Big)dt\leq$$
$$\frac{1}{4\delta_1}\|\eta^-_{q_x,m-1}\|^2+\delta_1\|\xi^+_{q_x,m-1}\|^2+\frac{1}{4\delta_2}\|\eta^-_{q_y,m-1}\|^2+\delta_2\|\xi^+_{q_y,m-1}\|^2$$
$$+\frac{1}{4\delta_3}\|\eta^-_{\rho,m-1}\|^2+\delta_3\|\xi^+_{\rho,m-1}\|^2+c_9\int_{I_m}(\|\xi_\rho\|^2+\|\xi_{q_x}\|^2+\|\xi_{q_y}\|^2)dt$$
$$+c_{10}\int_{I_m}(\|\eta_\rho\|^2+\|\eta_{q_x}\|^2+\|\eta_{q_y}\|^2+\|\eta_\sigma\|^2)dt+\int_{I_m}\Big((\frac{c_{11}}{h}+\frac{\beta}{2})\|\eta_{\lambda_{q_x}}-\eta_{q_x}\|^2_{L^2(\Gamma_{h,n})}$$
$$+(\frac{c_{12}}{h}+\frac{\alpha}{2})\|\eta_{\lambda_\rho}-\eta_\rho\|^2_{L^2(\Gamma_{h,n})}+(\frac{c_{13}}{h}+\frac{\beta'}{2})\|\eta_{\lambda_{q_y}}-\eta_{q_y}\|^2_{L^2(\Gamma_{h,n})}\Big)dt$$
$$+\frac{1}{4\delta_1}\|\xi^-_{q_{x,m-1}}\|^2+\delta_1\|\xi^+_{q_{x,m-1}}\|^2+\frac{1}{4\delta_1}\|\xi^-_{q_{y,m-1}}\|^2+\delta_1\|\xi^+_{q_{y,m-1}}\|^2 \tag{6.41}$$

This is the first part of the error estimate that we need for our further analysis. Before moving to the next part, we can write this estimate in another form. If we use (6.20) and (6.22) instead of (6.19) and (6.21) in our estimates, we would have

$$\frac{1}{2}(\|\xi^-_{\rho,m}\|^2 - \|\xi^-_{\rho,m-1}\|^2) + \frac{1}{4}\|[\![\xi]\!]_{\rho,m-1}\|^2 + \int_{I_m}\Big((\frac{\alpha}{2} - \frac{c_0}{h})\langle(\xi_\rho - \xi_{\lambda_\rho}),(\xi_\rho - \xi_{\lambda_\rho})\rangle\Big)dt$$
$$+ \frac{1}{2}(\|\xi^-_{q_x,m}\|^2 - \|\xi^-_{q_x,m-1}\|^2) + \frac{1}{4}\|[\![\xi]\!]_{q_x,m-1}\|^2$$
$$+ \int_{I_m}\Big((\frac{\beta}{2} - \frac{c_1}{h} - c_2)\langle(\xi_{q_x} - \xi_{\lambda_{q_x}}),(\xi_{q_x} - \xi_{\lambda_{q_x}})\rangle\Big)dt$$
$$+ \frac{1}{2}(\|\xi^-_{q_y,m}\|^2 - \|\xi^-_{q_y,m-1}\|^2) + \frac{1}{4}\|[\![\xi]\!]_{q_y,m-1}\|^2$$
$$+ \int_{I_m}\Big((\frac{\beta'}{2} - \frac{c_4}{h} - c_5)\langle(\xi_{q_y} - \xi_{\lambda_{q_y}}),(\xi_{q_y} - \xi_{\lambda_{q_y}})\rangle\Big)dt$$
$$+ \int_{I_m}\Big(c_7\epsilon_1\|\nabla\xi_{q_x}\|^2 + c_8\epsilon_2\|\nabla\xi_{q_y}\|^2 + (\frac{c_3}{\epsilon_1})\|\xi_{\sigma_x}\|^2 + (\frac{c_6}{\epsilon_2})\|\xi_{\sigma_y}\|^2\Big)dt \leq$$
$$\|\eta^-_{q_x,m-1}\|^2 + \|\eta^-_{q_y,m-1}\|^2 + \|\eta^-_{\rho,m-1}\|^2 + c_9\int_{I_m}(\|\xi_\rho\|^2 + \|\xi_{q_x}\|^2 + \|\xi_{q_y}\|^2)dt$$
$$+ c_{10}\int_{I_m}(\|\eta_\rho\|^2 + \|\eta_{q_x}\|^2 + \|\eta_{q_y}\|^2 + \|\eta_\sigma\|^2)dt + \int_{I_m}\Big((\frac{c_{11}}{h} + \frac{\beta}{2})\|\eta_{\lambda_{q_x}} - \eta_{q_x}\|^2_{L^2(\Gamma_{h,n})}$$
$$+ (\frac{c_{12}}{h} + \frac{\alpha}{2})\|\eta_{\lambda_\rho} - \eta_\rho\|^2_{L^2(\Gamma_{h,n})} + (\frac{c_{13}}{h} + \frac{\beta'}{2})\|\eta_{\lambda_{q_y}} - \eta_{q_y}\|^2_{L^2(\Gamma_{h,n})}\Big)dt. \qquad (6.42)$$

This form is more appropriate when we want to sum over all the time slabs as the first two terms on the left are in the form of telescoping sum.

The goal is now to bound the $\int_{I_m}(\|\xi_\rho\|^2 + \|\xi_{q_x}\|^2 + \|\xi_{q_y}\|^2)dt$ terms on the right hand side of (6.41) and (6.42).

## 6.4 Estimate of $\int_{I_m}(\|\xi_\rho\|^2 + \|\xi_{q_x}\|^2 + \|\xi_{q_y}\|^2)dt$

We need to introduce the discrete characteristic function as follows [14]:

**Theorem 6.4.1.** *Let $p \in \mathcal{P}^k(0,\tau;V_h)$, $\tau > 0$ and $t \in (0,\tau)$. We define the discrete characteristic function $\chi_{[0,t)}p$ as a polynomial $\tilde{p} \in \mathcal{P}^k(0,\tau;V_h)$ such*

*that*

$$\int_0^\tau \tilde{p}q = \int_0^t pq \quad \forall q \in \mathcal{P}^{k-1}(0,\tau;V_h)$$
$$\tilde{p}(0) = p(0). \tag{6.43}$$

*Moreover, there exist constants $C_k$ and $C$ such that*

$$\|\tilde{p}\|_{L^2(0,\tau;V)} \leq (1+C_k)\|p\|_{L^2(0,\tau;V)},$$
$$\|\tilde{p}-p\|_{L^2(0,\tau;V)} \leq C_k\|p\|_{L^2(0,\tau;V)},$$
$$\sum_K \int_0^\tau |\tilde{p}|^2_{H^1(K)}d\theta \leq C\sum_K \int_0^\tau |p|^2_{H^1(K)}d\theta. \tag{6.44}$$

*where $\mathcal{P}^k(0,\tau;V_h)$ as defined before, is the Bochner space of functions of polynomial of order $k$ over the interval $(0,\tau)$ with values in a finite dimensional subspace $V_h$ of a Banach space $V$*

Inequalities in (6.44) are essentially the $L^2$ and $H^1$ stability of the defined projection. It can also be proved that the discrete characteristic function is translationally invariant (see [34]). This means that we can shift the lower bound (0) and upper bound (t) of the integral in (6.43) to $t_{m-1}$ and $t_m$ respectively. We define

$$t_{m-1+\frac{l}{r}} = t_{m-1} + \frac{l}{r}(t_m - t_{m-1}), \quad l = 0, ..., r.$$

where $r$ is the degree of polynomial in time. We are now going to use $\tilde{\xi}_\rho$ as a test function in equation (6.17). Based on the definition of discrete characteristic

function, the first two terms can be simplified to obtain

$$\int_{I_m}(\frac{\partial\xi_\rho}{\partial t},\tilde{\xi}_\rho)dt+((\xi^+_{\rho,m-1}-\xi^-_{\rho,m-1}),\tilde{\xi}^+_{\rho,m-1})=$$
$$\int_{t_{m-1}}^{t_{m-1+\frac{l}{r}}}(\frac{\partial\xi_\rho}{\partial t},\xi_\rho)dt+((\xi^+_{\rho,m-1}-\xi^-_{\rho,m-1}),\xi^+_{\rho,m-1}),$$

where $t=t_{m-1+\frac{l}{r}}$ in (6.43) and we have used the equality $\tilde{\xi}^+_{\rho,m-1}=\xi^+_{\rho,m-1}$. Therefore

$$\int_{I_m}(\frac{\partial\xi_\rho}{\partial t},\tilde{\xi}_\rho)dt+((\xi^+_{\rho,m-1}-\xi^-_{\rho,m-1}),\tilde{\xi}^+_{\rho,m-1})=\frac{1}{2}(\|\xi_{\rho,m-1+\frac{l}{r}}\|^2+\|\xi^+_{\rho,m-1}\|^2)-(\xi^-_{\rho,m-1},\xi^+_{\rho,m-1})$$

From the linearity of the projection (6.43), choosing $-\tilde{\xi}_{\lambda_\rho}$ as a test function in (6.12) and adding it to the result, (see (6.17)) we will have

$$\frac{1}{2}(\|\xi_{\rho,m-1+\frac{l}{r}}\|^2+\|\xi^+_{\rho,m-1}\|^2)+\int_{I_m}\Big(\alpha\langle(\xi_\rho-\xi_{\lambda_\rho}),\widetilde{\xi_\rho-\xi_{\lambda_\rho}}\rangle\Big)dt=$$
$$\int_{I_m}\Big(\frac{1}{\epsilon_1}(\xi_{\sigma_{x1}},\tilde{\xi}_\rho)+\frac{1}{\epsilon_2}(\xi_{\sigma_{y2}},\tilde{\xi}_\rho)+\frac{1}{\epsilon_1}(\eta_{\sigma_{x1}},\tilde{\xi}_\rho)+\frac{1}{\epsilon_2}(\eta_{\sigma_{y2}},\tilde{\xi}_\rho)$$
$$+\alpha\langle(\eta_\rho-\eta_{\lambda_\rho}),\widetilde{\xi_\rho-\xi_{\lambda_\rho}}\rangle\Big)dt+(\xi^-_{\rho,m-1},\xi^+_{\rho,m-1})$$

Using Cauchy inequality on $\alpha\langle(\eta_\rho-\eta_{\lambda_\rho}),\widetilde{\xi_\rho-\xi_{\lambda_\rho}}\rangle$ term on the right hand side and bringing the tilde term to the left, the integral equation of the left can be written as

$$\int_{I_m}\Big(\alpha\langle(\xi_\rho-\xi_{\lambda_\rho}),\widetilde{\xi_\rho-\xi_{\lambda_\rho}}\rangle-\frac{\alpha}{2}\|\widetilde{\xi_\rho-\xi_{\lambda_\rho}}\|^2\Big)dt$$

which can be simplified as follows. Taking $\xi_\rho-\xi_{\lambda_\rho}=p$ for simplicity, we have:

$$\alpha\langle p,\tilde{p}\rangle-\frac{\alpha}{2}\langle\tilde{p},\tilde{p}\rangle=\frac{\alpha}{2}\langle p,\tilde{p}\rangle+\frac{\alpha}{2}\langle p,\tilde{p}\rangle-\frac{\alpha}{2}\langle\tilde{p},\tilde{p}\rangle=\frac{\alpha}{2}\langle p,\tilde{p}\rangle+\frac{\alpha}{2}\langle p-\tilde{p},\tilde{p}\rangle=\frac{\alpha}{2}\langle p,\tilde{p}\rangle$$
$$+\frac{\alpha}{2}\langle p-\tilde{p},\tilde{p}-p\rangle+\frac{\alpha}{2}\langle p-\tilde{p},p\rangle=\frac{\alpha}{2}\langle p,p\rangle-\frac{\alpha}{2}\langle p-\tilde{p},p-\tilde{p}\rangle$$

Using the bound in (6.44), we have

$$\frac{\alpha}{2}\langle p,p\rangle - \frac{\alpha}{2}\langle p-\tilde{p},p-\tilde{p}\rangle \;>\; (1-C_k)\frac{\alpha}{2}\langle p,p\rangle$$

which in general is a negative term. We keep this negative term for now on the left hand side. Therefore we obtain

$$\begin{aligned}
&\|\xi_{\rho,m-1+\frac{l}{r}}\|^2 + \|\xi^+_{\rho,m-1}\|^2 + \int_{I_m}(1-C_k)\frac{\alpha}{2}\|(\xi_\rho-\xi_{\lambda_\rho})\|^2_{L^2(\Gamma_{h,n})} \leq \\
&+\int_{I_m}\Big(c_1\|\xi_\rho\|^2 + \frac{\alpha}{2}\|\eta_\rho-\eta_{\lambda_\rho}\|^2_{L^2(\Gamma_{h,n})} + \frac{c_2}{\epsilon_1}\|\xi_{\sigma_x}\|^2 + \frac{c_2}{\epsilon_2}\|\xi_{\sigma_y}\|^2 \\
&+\frac{c_3}{\epsilon_1}\|\eta_{\sigma_x}\|^2 + \frac{c_3}{\epsilon_2}\|\eta_{\sigma_y}\|^2\Big)dt \\
&+\frac{1}{4\delta_1}\|(\eta^-_{\rho,m-1})\|^2 + \delta_1\|\xi^+_{\rho,m-1}\|^2 + \frac{1}{4\delta_2}\|(\xi^-_{\rho,m-1})\|^2 + \delta_2\|\xi^+_{\rho,m-1}\|^2. \qquad (6.45)
\end{aligned}$$

We can now sum over $l = 1, ..., r-1$ and dividing by $(r-1)$ to obtain

$$\begin{aligned}
&\frac{1}{(r-1)}\sum_{l=1}^{r-1}\|\xi_{\rho,m-1+\frac{l}{r}}\|^2 + \|\xi^+_{\rho,m-1}\|^2 + \int_{I_m}(1-C_k)\frac{\alpha}{2}\|(\xi_\rho-\xi_{\lambda_\rho})\|^2_{L^2(\Gamma_{h,n})} \leq \\
&+\int_{I_m}\Big(c_1\|\xi_\rho\|^2 + \frac{\alpha}{2}\|\eta_\rho-\eta_{\lambda_\rho}\|^2_{L^2(\Gamma_{h,n})} + \frac{c_2}{\epsilon_1}\|\xi_{\sigma_x}\|^2 + \frac{c_2}{\epsilon_2}\|\xi_{\sigma_y}\|^2 \\
&+\frac{c_3}{\epsilon_1}\|\eta_{\sigma_x}\|^2 + \frac{c_3}{\epsilon_2}\|\eta_{\sigma_y}\|^2\Big)dt \\
&+\frac{1}{4\delta_1}\|(\eta^-_{\rho,m-1})\|^2 + \delta_1\|\xi^+_{\rho,m-1}\|^2 + \frac{1}{4\delta_2}\|(\xi^-_{\rho,m-1})\|^2 + \delta_2\|\xi^+_{\rho,m-1}\|^2 \qquad (6.46)
\end{aligned}$$

Going through a similar procedure for $q_x$, i.e. using $\tilde{\xi}_{q_x}$ and $\tilde{\xi}_{\lambda_{q_x}}$ as test

functions correspondingly we obtain

$$\int_{I_m}(\frac{\partial \xi_{q_x}}{\partial t},\tilde{\xi}_{q_x})dt+((\xi^+_{q_{x_{m-1}}}-\xi^-_{q_{x_{m-1}}}),\tilde{\xi}^+_{q_{x_{m-1}}})+\int_{I_m}\Big(\langle\xi_{\sigma_x}\cdot n,\tilde{\xi}_{q_x}-\tilde{\xi}_{\lambda_{q_x}}\rangle$$
$$-(\xi_{\sigma_x},\nabla\tilde{\xi}_{q_x})+\beta\langle(\xi_{q_x}-\xi_{\lambda_{q_x}}),\tilde{\xi}_{q_x}-\tilde{\xi}_{\lambda_{q_x}}\rangle\Big)dt=$$
$$\int_{I_m}(\frac{\partial \eta_{q_x}}{\partial t},\tilde{\xi}_{q_x})dt+((\eta^+_{q_{x_{m-1}}}-\eta^-_{q_{x_{m-1}}}),\tilde{\xi}^+_{q_{x_{m-1}}})+\int_{I_m}\Big(\langle\eta_{\sigma_x}\cdot n,\tilde{\xi}_{q_x}-\tilde{\xi}_{\lambda_{q_x}}\rangle$$
$$-(\eta_{\sigma_x},\nabla\tilde{\xi}_{q_x})-\langle\frac{\lambda^2_{q_x}}{\lambda_\rho}+g\frac{\lambda^2_\rho}{2},\tilde{\xi}_{q_x}n_x\rangle+\langle\frac{\lambda^2_{q_{x_{h\tau}}}}{\lambda_{\rho_{h\tau}}}+g\frac{\lambda^2_{\rho_{h\tau}}}{2},\tilde{\xi}_{q_x}n_x\rangle+(\frac{{q_x}^2}{\rho}+\frac{g\rho^2}{2},\partial_x\tilde{\xi}_{q_x})$$
$$-(\frac{{q_{x_{h\tau}}}^2}{\rho_{h\tau}}+\frac{g\rho^2_{h\tau}}{2},\partial_x\tilde{\xi}_{q_x})-\langle\frac{\lambda_{q_y}\lambda_{q_x}}{\lambda_\rho},\tilde{\xi}_{q_x}n_y\rangle+\langle\frac{\lambda_{q_{y_{h\tau}}}\lambda_{q_{x_{h\tau}}}}{\lambda_{\rho_{h\tau}}},\tilde{\xi}_{q_x}n_y\rangle$$
$$+(\frac{q_xq_y}{\rho},\partial_y\tilde{\xi}_{q_x})-(\frac{q_{x_{h\tau}}q_{y_{h\tau}}}{\rho_{h\tau}},\partial_y\tilde{\xi}_{q_x})+\beta\langle(\eta_{q_x}-\eta_{\lambda_{q_x}}),\tilde{\xi}_{q_x}-\tilde{\xi}_{\lambda_{q_x}}\rangle\Big)dt.$$

Different terms on the left or right hand side can be bounded as before, e.g.

$$\begin{aligned}(\xi_{\sigma_x},\nabla\tilde{\xi}_{q_x})&\leq\frac{1}{c_1\epsilon_1}\|\xi_{\sigma_x}\|^2+c_1\epsilon_1\|\nabla\tilde{\xi}_{q_x}\|^2\\&\leq\frac{1}{c_1\epsilon_1}\|\xi_{\sigma_x}\|^2+c_2\epsilon_1\|\nabla\xi_{q_x}\|^2.\end{aligned}$$

$c_1$ can be taken large enough so that we can absorb the first term (on the left hand side) later on. We are also going to use the Lipschitz continuity of the advective flux as before and bound the terms on the right hand side using (6.44), as an example

$$\begin{aligned}\langle|\xi_{\lambda_\rho}||n_x|,|\tilde{\xi_{q_x}}|\rangle\leq\langle|\xi_{\lambda_\rho}|,|\tilde{\xi_{q_x}}|\rangle&=\langle|\xi_{\lambda_\rho}|-|\xi_\rho|,|\tilde{\xi_{q_x}}|\rangle+\langle|\xi_\rho|,|\tilde{\xi_{q_x}}|\rangle\\&\leq\langle|\xi_{\lambda_\rho}-\xi_\rho|,|\tilde{\xi_{q_x}}|\rangle+\langle|\xi_\rho|,|\tilde{\xi_{q_x}}|\rangle\\&\leq\frac{c_1}{h}\|\xi_{\lambda_\rho}-\xi_\rho\|^2_{L^2(\Gamma_{h,n})}+c_2\|\xi_{q_x}\|^2+\frac{c_3}{\epsilon_{1\circ}}\|\xi_\rho\|^2\\&\quad+c_4\epsilon_1\|\nabla\xi_{q_x}\|^2\end{aligned}$$

If it's assumed that the space of trace of $\xi_{q_x}$ is a subset of the space of $\xi_{\lambda_{q_x}}$ (a compatibility condition, see assumption (iii)), then from the linearity of the projection (6.43), and using (6.44), we will have

$$\|\tilde{\xi_q} - \tilde{\xi_{\lambda_q}}\|^2_{L^2(\Gamma_{h,n})} = \|\widetilde{(\xi_q - \xi_{\lambda_q})}\|^2_{L^2(\Gamma_{h,n})} \leq C\|\xi_q - \xi_{\lambda_q}\|^2_{L^2(\Gamma_{h,n})}.$$

Thus after summing over $l = 1, ..., r-1$ and dividing by $(r-1)$ we will obtain

$$\begin{aligned}
&\frac{1}{(r-1)}\sum_{l=1}^{r-1}\|\xi_{q_{x,m-1+\frac{l}{r}}}\|^2 + \|\xi^+_{q_x,m-1}\|^2 \\
&+ \int_{I_m}\Big(((1-C_k)\frac{\beta}{2} - \frac{c_0}{h} - c_7)\langle(\xi_{q_x} - \xi_{\lambda_{q_x}}), (\xi_{q_x} - \xi_{\lambda_{q_x}})\rangle\Big)dt \\
&\leq c_1\int_{I_m}(\|\xi_\rho\|^2 + \|\xi_{q_x}\|^2 + \|\xi_{q_y}\|^2)dt + c_2\int_{I_m}(\|\eta_\rho\|^2 + \|\eta_{q_x}\|^2 + \|\eta_{q_y}\|^2 + \|\eta_{\sigma_x}\|^2)dt \\
&+ \int_{I_m}\Big(\frac{c_3}{h}\|\eta_{\lambda_{q_x}} - \eta_{q_x}\|^2_{L^2(\Gamma_{h,n})} + \frac{c_4}{h}\|\eta_{\lambda_\rho} - \eta_\rho\|^2_{L^2(\Gamma_{h,n})} + c\epsilon_1\|\nabla\xi_{q_x}\|^2 + c\epsilon_2\|\nabla\xi_{q_y}\|^2 \\
&+ \frac{c_{10}}{\epsilon_1}\|\xi_{\sigma_x}\|^2 + \frac{\beta}{2}\|\eta_{q_x} - \eta_{\lambda_{q_x}}\|^2_{L^2(\Gamma_{h,n})} + \frac{c_8}{h}\|\eta_{\lambda_{q_y}} - \eta_{q_y}\|^2_{L^2(\Gamma_{h,n})} \\
&+ \frac{c_9}{h}\|\xi_{\lambda_{q_y}} - \xi_{q_y}\|^2 + \frac{c_{10}}{h}\|\xi_{\lambda_\rho} - \xi_\rho\|^2\Big)dt + \frac{1}{4\delta_1}\|(\eta^-_{q_x,m-1})\|^2 + \delta_1\|\xi^+_{q_x,m-1}\|^2 \\
&+ \frac{1}{4\delta_2}\|(\xi^-_{q_x,m-1})\|^2 + \delta_2\|\xi^+_{q_x,m-1}\|^2 \qquad (6.47)
\end{aligned}$$

Adding (6.46) and (6.47) and the error inequality corresponding to $q_y$, we

obtain

$$\frac{1}{(r-1)}\sum_{l=1}^{r-1}\|\xi_{\rho,m-1+\frac{l}{r}}\|^2+\|\xi^+_{\rho,m-1}\|^2+\int_{I_m}(1-C_k)\frac{\alpha}{2}\|(\xi_\rho-\xi_{\lambda_\rho})\|^2_{L^2(\Gamma_{h,n})}dt$$
$$+\frac{1}{(r-1)}\sum_{l=1}^{r-1}\|\xi_{q_x,m-1+\frac{l}{r}}\|^2+\|\xi^+_{q_x,m-1}\|^2$$
$$+\int_{I_m}((1-C_k)\frac{\beta}{2}-\frac{c_0}{h}-c_7)\|(\xi_{q_x}-\xi_{\lambda_{q_x}})\|^2_{L^2(\Gamma_{h,n})}dt$$
$$+\frac{1}{(r-1)}\sum_{l=1}^{r-1}\|\xi_{q_y,m-1+\frac{l}{r}}\|^2+\|\xi^+_{q_y,m-1}\|^2$$
$$+\int_{I_m}((1-C_k)\frac{\beta}{2}-\frac{c_0}{h}-c_7)\|(\xi_{q_y}-\xi_{\lambda_{q_y}})\|^2_{L^2(\Gamma_{h,n})}dt$$
$$\leq c_1\int_{I_m}(\|\xi_\rho\|^2+\|\xi_{q_x}\|^2+\|\xi_{q_y}\|^2)dt+c_2\int_{I_m}(\|\eta_\rho\|^2+\|\eta_{q_x}\|^2+\|\eta_{q_y}\|^2+\|\eta_{\sigma_x}\|^2)dt$$
$$+\int_{I_m}\Big(\frac{c_3}{h}\|\eta_{\lambda_{q_x}}-\eta_{q_x}\|^2_{L^2(\Gamma_{h,n})}+\frac{c_4}{h}\|\eta_{\lambda_{q_y}}-\eta_{q_y}\|^2_{L^2(\Gamma_{h,n})}+\frac{c_5}{h}\|\eta_{\lambda_\rho}-\eta_\rho\|^2_{L^2(\Gamma_{h,n})}$$
$$+c_5\epsilon\|\nabla\xi_{q_x}\|^2+c_5\epsilon\|\nabla\xi_{q_y}\|^2+(\frac{c_6}{\epsilon})\|\xi_{\sigma_x}\|^2+(\frac{c_6}{\epsilon})\|\xi_{\sigma_y}\|^2$$
$$+\frac{\beta}{2}\|\eta_{q_x}-\eta_{\lambda_{q_x}}\|^2_{L^2(\Gamma_{h,n})}+\frac{\beta'}{2}\|\eta_{q_y}-\eta_{\lambda_{q_y}}\|^2_{L^2(\Gamma_{h,n})}+\frac{c_7}{\epsilon_1}\|\eta_{\sigma_x}\|^2+\frac{c_7}{\epsilon_2}\|\eta_{\sigma_y}\|^2\Big)dt$$
$$+\frac{1}{4\delta_1}\|(\eta^-_{\rho,m-1})\|^2+\delta_1\|\xi^+_{\rho,m-1}\|^2+\frac{1}{4\delta_2}\|(\xi^-_{\rho,m-1})\|^2+\delta_2\|\xi^+_{\rho,m-1}\|^2$$
$$+\frac{1}{4\delta_3}\|(\eta^-_{q_x,m-1})\|^2+\delta_3\|\xi^+_{q_x,m-1}\|^2+\frac{1}{4\delta_4}\|(\xi^-_{q_x,m-1})\|^2+\delta_4\|\xi^+_{q_x,m-1}\|^2$$
$$+\frac{1}{4\delta_5}\|(\eta^-_{q_y,m-1})\|^2+\delta_5\|\xi^+_{q_y,m-1}\|^2+\frac{1}{4\delta_6}\|(\xi^-_{q_y,m-1})\|^2+\delta_6\|\xi^+_{q_y,m-1}\|^2. \qquad (6.48)$$

As stated before the penalty terms on the left hand side of (6.48) are in general not positive, also we need to bound the terms like $\|\xi_{\sigma_y}\|^2$ and $\|\nabla\xi_{q_x}\|^2$ and other similar terms on the right hand side too. Therefore we are going to add a multiple of equation (6.41) to it. Choosing $C = 2\times max(C_K, c_5, c_6)$,

multiplying (6.41) by $C$ and adding it to (6.48), we obtain

$$\tilde{C}_1(\|\xi^-_{\rho,m}\|^2+\sum_{l=1}^{r-1}\|\xi_{\rho,m-1+\frac{l}{r}}\|^2+\|\xi^+_{\rho,m-1}\|^2)+\tilde{C}_2(\|\xi^-_{q_x,m}\|^2$$
$$+\sum_{l=1}^{r-1}\|\xi_{q_x,m-1+\frac{l}{r}}\|^2+\|\xi^+_{q_x,m-1}\|^2)+\tilde{C}_3(\|\xi^-_{q_y,m}\|^2+\sum_{l=1}^{r-1}\|\xi_{q_y,m-1+\frac{l}{r}}\|^2+\|\xi^+_{q_y,m-1}\|^2)$$
$$+\int_{I_m}(c_1\alpha-c_2-\frac{c_3}{h})\langle(\xi_\rho-\xi_{\lambda_\rho}),(\xi_\rho-\xi_{\lambda_\rho})\rangle$$
$$+(c_4\beta-c_5-\frac{c_6}{h})\langle(\xi_{q_x}-\xi_{\lambda_{q_x}}),(\xi_{q_x}-\xi_{\lambda_{q_x}})\rangle$$
$$+(c_7\beta'-c_8-\frac{c_9}{h})\langle(\xi_{q_y}-\xi_{\lambda_{q_y}}),(\xi_{q_y}-\xi_{\lambda_{q_y}})\rangle$$
$$+\frac{c_{10}}{\epsilon_1}\|\xi_{\sigma_x}\|^2+\frac{c_{11}}{\epsilon_2}\|\xi_{\sigma_y}\|^2+c_{12}\epsilon_1\|\nabla\xi_{q_x}\|^2+c_{13}\epsilon_2\|\nabla\xi_{q_y}\|^2\leq$$
$$+\int_{I_m}C_0\Big(\|\xi_\rho\|^2+\|\xi_{q_x}\|^2+\|\xi_{q_y}\|^2\Big)dt+\int_{I_m}\frac{C_2}{h}\Big(\|\eta_\rho-\eta_{\lambda_\rho}\|^2_{L^2(\Gamma_{h,n})}$$
$$+\|\eta_{q_x}-\eta_{\lambda_{q_x}}\|^2_{L^2(\Gamma_{h,n})}+\|\eta_{q_y}-\eta_{\lambda_{q_y}}\|^2_{L^2(\Gamma_{h,n})}\Big)dt+\int_{I_m}\|\eta_\sigma\|^2_{H^1}dt$$
$$+C_3\Big(\|\eta^-_{\rho,m-1}\|^2+\|\eta^-_{q_x,m-1}\|^2+\|\eta^-_{q_y,m-1}\|^2+\|(\xi^-_{\rho,m-1})\|^2+\|(\xi^-_{q_x,m-1})\|^2$$
$$+\|(\xi^-_{q_y,m-1})\|^2+\delta\|(\xi^+_{\rho,m-1})\|^2+\delta\|(\xi^+_{q_x,m-1})\|^2+\delta\|(\xi^+_{q_y,m-1})\|^2\Big).\tag{6.49}$$

where the sum on the left hand side can be simplified as

$$\|\xi^-_{\rho,m}\|^2+\sum_{l=1}^{r-1}\|\xi_{\rho,m-1+\frac{l}{r}}\|^2+\|\xi^+_{\rho,m-1}\|^2=\sum_{l=0}^{r}\|\xi_{\rho,m-1+\frac{l}{r}}\|^2$$

and similarly for $\xi_{q_x},\xi_{q_y}$. As can be seen from (6.49), in order to have a positive norms on the left we should have, $\alpha=\mathcal{O}(\frac{1}{h}),\beta=\mathcal{O}(\frac{1}{h}),\beta'=\mathcal{O}(\frac{1}{h})$. Now using the equivalence of norms in finite dimensional space [11], i.e.

$$\sum_{l=0}^{r}\|\xi_{m-1+\frac{l}{r}}\|^2_{L^2(\Omega)}\geq\frac{L_r}{\tau_m}\int_{I_m}\|\xi\|^2_{L^2(\Omega)}dt\tag{6.50}$$

$$\|\xi^+_{m-1}\|^2_{L^2(\Omega)}\leq\frac{M_r}{\tau_m}\int_{I_m}\|\xi\|^2_{L^2(\Omega)}dt\tag{6.51}$$

where $\xi$ can be $\xi_\rho, \xi_{q_x}, \xi_{q_y}$, and ignoring the other positive terms on the left for now, (6.49) can be written as

$$(\frac{L_r\tilde{C}}{\tau_m} - \frac{\delta C_3 M_r}{\tau_m} - C_0)\int_{I_m}(\|\xi_\rho\|^2 + \|\xi_{q_x}\|^2 + \|\xi_{q_y}\|^2)dt \leq \int_{I_m} C_1\Big(\|\eta_\rho\|^2_{H^1} + \|\eta_{q_x}\|^2_{H^1}$$
$$+ \|\eta_{q_y}\|^2_{H^1} + \|\eta_\sigma\|^2_{H^1} + \|\eta_{\lambda_{q_x}}\|_{H^1} + \|\eta_{\lambda_{q_y}}\|_{H^1}\Big)dt + C_2\Big(\|\eta^-_{\rho,m-1}\|^2 + \|\eta^-_{q_x,m-1}\|^2$$
$$+ \|\eta^-_{q_y,m-1}\|^2 + \|(\xi^-_{\rho,m-1})\|^2 + \|(\xi^-_{q_x,m-1})\|^2 + \|(\xi^-_{q_y,m-1})\|^2\Big)dt. \tag{6.52}$$

Choosing

$$\delta = \frac{L_r\tilde{C}}{2C_3M_r},$$

the condition

$$\frac{L_r\tilde{C}}{2\tau_m} - C_0 > 0$$

is satisfied by choosing $\tau_m$ small enough, i.e.

$$\tau_m < \frac{L_r\tilde{C}}{2C_0}. \tag{6.53}$$

Therefore (6.52) can be written as

$$\int_{I_m}(\|\xi_\rho\|^2 + \|\xi_{q_x}\|^2 + \|\xi_{q_y}\|^2)dt \leq \tau_m\int_{I_m} C_1\Big(\|\eta_\rho\|^2_{H^1} + \|\eta_{q_x}\|^2_{H^1} + \|\eta_{q_y}\|^2_{H^1}$$
$$+ \|\eta_\sigma\|^2_{H^1} + \|\eta_{\lambda_\rho}\|^2_{H^1} + \|\eta_{\lambda_{q_x}}\|_{H^1} + \|\eta_{\lambda_{q_y}}\|_{H^1}\Big)dt + \tau_m C_2\Big(\|\eta^-_{\rho,m-1}\|^2 + \|\eta^-_{q_x,m-1}\|^2$$
$$+ \|\eta^-_{q_y,m-1}\|^2 + \|(\xi^-_{\rho,m-1})\|^2 + \|(\xi^-_{q_x,m-1})\|^2 + \|(\xi^-_{q_y,m-1})\|^2\Big)dt \tag{6.54}$$

which is the $L^2$ bound we were looking for. Using this bound in (6.42) and take the sum from $m = 1, ..., M$, where $M$ is the total number of time slabs

and noticing that $\|\xi^-_{\rho,0}\|^2 = \|\xi^-_{q_x,0}\|^2 = \|\xi^-_{q_y,0}\|^2 = 0$, we will have

$$
\begin{aligned}
&\|\xi^-_{\rho,M}\|^2 + \|\xi^-_{q_x,M}\|^2 + \|\xi^-_{q_y,M}\|^2 + \sum_{m=1}^{M} \int_{I_m} \Big( \alpha \|\xi_\rho - \xi_{\lambda_\rho})\|^2_{L^2(\Gamma_{h,n})} \\
&+ \beta \|(\xi_{q_x} - \xi_{\lambda_{q_x}})\|^2_{L^2(\Gamma_{h,n})} + \beta' \|(\xi_{q_y} - \xi_{\lambda_{q_y}})\|^2_{L^2(\Gamma_{h,n})} \Big) dt \\
&+ \sum_{m=1}^{M} \int_{I_m} \Big( \frac{c_1}{\epsilon_1} \|\xi_{\sigma_x}\|^2 + \frac{c_2}{\epsilon_2} \|\xi_{\sigma_y}\|^2 + c_3 \epsilon_1 \|\nabla \xi_{q_x}\|^2 + c_4 \epsilon_2 \|\nabla \xi_{q_y}\|^2 \Big) dt \\
&\leq C \sum_{m=0}^{M-1} \tau_{m+1} \Big( \|(\xi^-_{\rho,m})\|^2 + \|(\xi^-_{q_x,m})\|^2 + \|(\xi^-_{q_y,m})\|^2 \Big) \\
&+ C_1 \sum_{m=1}^{M} (1 + \tau_m) \Big( \|\eta^-_{\rho,m-1}\|^2 + \|\eta^-_{q_x,m-1}\|^2 + \|\eta^-_{q_y,m-1}\|^2 \Big) \\
&+ C_2 \sum_{m=1}^{M} (1 + \tau_m) \int_{I_m} \Big( \|\eta_\rho\|^2_{H^1} + \|\eta_q\|^2_{H^1} + \|\eta_\sigma\|^2_{H^1} + \|\eta_{\lambda_\rho}\|^2_{H^1} + \|\eta_{\lambda_{q_x}}\|^2_{H^1} + \|\eta_{\lambda_{q_y}}\|^2_{H^1} \Big) dt,
\end{aligned}
$$

where we have used

$$
\begin{aligned}
&\sum_{m=1}^{M} \tau_m (\|(\xi^-_{\rho,m-1})\|^2 + \|(\xi^-_{q_x,m-1})\|^2 + \|(\xi^-_{q_y,m-1})\|^2) = \\
&\sum_{m=0}^{M-1} \tau_{m+1} (\|(\xi^-_{\rho,m})\|^2 + \|(\xi^-_{q_x,m})\|^2 + \|(\xi^-_{q_y,m})\|^2)
\end{aligned}
$$

and that $\alpha = \mathcal{O}(\frac{1}{h}), \beta = \mathcal{O}(\frac{1}{h}), \beta' = \mathcal{O}(\frac{1}{h})$.

We now need the following discrete Gronwall inequality:

Let $x_M, b_M, c_M \geq 0$ and $a_M > 0$ for $M = 0, 1, 2, ...$ and let the sequence $a_M$ be non-decreasing. If

$$
\begin{aligned}
&x_0 + c_0 \leq a_0 \\
&x_M + c_M \leq a_M + \sum_{j=0}^{M-1} b_j x_j
\end{aligned}
$$

Then

$$x_M + c_M \leq a_M \prod_{j=0}^{M-1}(1+b_j)$$

For our case

$$x_M = \|\xi^-_{\rho,M}\|^2 + \|\xi^-_{q_x,M}\|^2 + \|\xi^-_{q_y,M}\|^2$$

$$c_M = \sum_{m=1}^{M}\int_{I_m}\Big(\alpha\|\xi_\rho - \xi_{\lambda_\rho})\|_{L^2(\Gamma_{h,n})} + \beta\|(\xi_{q_x} - \xi_{\lambda_{q_x}})\|_{L^2(\Gamma_{h,n})}$$
$$+ \beta'\|(\xi_{q_y} - \xi_{\lambda_{q_y}})\|_{L^2(\Gamma_{h,n})}\Big)dt$$
$$+ \sum_{m=1}^{M}\int_{I_m}\Big(\frac{c_1}{\epsilon_1}\|\xi_{\sigma_x}\|^2 + \frac{c_2}{\epsilon_2}\|\xi_{\sigma_y}\|^2 + c_3\epsilon_1\|\nabla q_x\|^2 + c_4\epsilon_2\|\nabla q_y\|^2\Big)dt$$

$$a_M = C_1\sum_{m=1}^{M}(1+\tau_m)\Big(\|\eta^-_{\rho,m-1}\|^2 + \|\eta^-_{q_x,m-1}\|^2 + \|\eta^-_{q_y,m-1}\|^2\Big)$$
$$+ C_2\sum_{m=1}^{M}(1+\tau_m)\int_{I_m}\Big(\|\eta_\rho\|^2_{H^1} + \|\eta_q\|^2_{H^1} + \|\eta_\sigma\|^2_{H^1} + \|\eta_{\lambda_\rho}\|^2_{H^1} + \|\eta_{\lambda_{q_x}}\|^2_{H^1} + \|\eta_{\lambda_{q_y}}\|^2_{H^1}\Big)dt$$

$$b_j = C\tau_{j+1}.$$

The product term can be simplified as

$$\prod_{j=0}^{M-1}(1+b_j) = \prod_{j=0}^{M-1}(1+C\tau_{j+1}) = \prod_{j=1}^{M}(1+C\tau_j) \leq exp(C\sum_{j=1}^{M}\tau_j) = exp(CT),$$

where $T$ is the final time. Hence by substitution we obtain the following bound

$$\|\xi^-_{\rho,M}\|^2 + \|\xi^-_{q_x,M}\|^2 + \|\xi^-_{q_y,M}\|^2 + \sum_{m=1}^{M}\int_{I_m}\Big(\alpha\|\xi_\rho - \xi_{\lambda_\rho})\|^2_{L^2(\Gamma_{h,n})} + \beta\|(\xi_{q_x} - \xi_{\lambda_{q_x}})\|^2_{L^2(\Gamma_{h,n})}$$
$$+ \beta'\|(\xi_{q_y} - \xi_{\lambda_{q_y}})\|^2_{L^2(\Gamma_{h,n})}\Big)dt + \sum_{m=1}^{M}\int_{I_m}\Big(\frac{c_1}{\epsilon_1}\|\xi_{\sigma_x}\|^2 + \frac{c_2}{\epsilon_2}\|\xi_{\sigma_y}\|^2 + c_3\epsilon_1\|\nabla\xi_{q_x}\|^2$$
$$+ c_4\epsilon_2\|\nabla\xi_{q_y}\|^2\Big)dt \leq exp(CT)\Bigg(C_1\sum_{m=1}^{M}(1+\tau_m)\big(\|\eta^-_{\rho,m-1}\|^2 + \|\eta^-_{q_x,m-1}\|^2 + \|\eta^-_{q_y,m-1}\|^2\big)$$
$$+ C_2\sum_{m=1}^{M}(1+\tau_m)\int_{I_m}\Big(\|\eta_\rho\|^2_{H^1} + \|\eta_q\|^2_{H^1} + \|\eta_\sigma\|^2_{H^1} + \|\eta_{\lambda_\rho}\|^2_{H^1} + \|\eta_{\lambda_{q_x}}\|^2_{H^1} + \|\eta_{\lambda_{q_y}}\|^2_{H^1}\Big)\Bigg)dt. \tag{6.55}$$

This is the final form of error estimate that is obtained. We can extract the needed error estimate from this inequality. For example we can obtain the pointwise in time error at the end of the time step from (6.55), using

$$\|e^-_m\|^2 \leq 2(\|\xi^-_m\|^2 + \|\eta^-_m\|^2),$$

and neglecting the positive terms on left we obtain

$$\|e^-_{\rho,M}\|^2 + \|e^-_{q_x,M}\|^2 + \|e^-_{q_y,M}\|^2 \leq exp(CT)\Bigg(C_1\sum_{m=1}^{M}(1+\tau_m)\big(\|\eta^-_{\rho,m-1}\|^2 + \|\eta^-_{q_x,m-1}\|^2$$
$$+ \|\eta^-_{q_y,m-1}\|^2\big) + C_2\sum_{m=1}^{M}(1+\tau_m)\int_{I_m}\Big(\|\eta_\rho\|^2_{H^1} + \|\eta_q\|^2_{H^1} + \|\eta_\sigma\|^2_{H^1} + \|\eta_{\lambda_\rho}\|^2_{H^1} + \|\eta_{\lambda_{q_x}}\|^2_{H^1}$$
$$+ \|\eta_{\lambda_{q_y}}\|^2_{H^1}\Big)\Bigg)dt.$$

For $L^2$ norm we can use (6.54). After summing over time slabs we have

$$\int_0^T (\|e_\rho\|^2 + \|e_{q_x}\|^2 + \|e_{q_y}\|^2)dt \leq \sum_{m=1}^{M} \tau_m \int_{I_m} C_1\Big(\|\eta_\rho\|^2_{H^1} + \|\eta_{q_x}\|^2_{H^1} + \|\eta_{q_y}\|^2_{H^1}$$
$$+ \|\eta_\sigma\|^2_{H^1} + \|\eta_{\lambda_\rho}\|^2_{H^1} + \|\eta_{\lambda_{q_x}}\|_{H^1} + \|\eta_{\lambda_{q_y}}\|_{H^1}\Big)dt + \sum_{m=1}^{M} \tau_m C_2\Big(\|\eta^-_{\rho,m-1}\|^2 + \|\eta^-_{q_x,m-1}\|^2$$
$$+ \|\eta^-_{q_y,m-1}\|^2 + \|(\xi^-_{\rho,m-1})\|^2 + \|(\xi^-_{q_x,m-1})\|^2 + \|(\xi^-_{q_y,m-1})\|^2\Big)dt \tag{6.56}$$

We can now use the properties of the space-time projection defined in (6.5) and (6.6). The projection is split into time and space, i.e.:

$$\eta|_{I_m} = (\pi u - u)|_{I_m} = u - \pi_h u + \pi_h u - \pi_t(\pi_h)u = \eta^{(1)}|_{I_m} + \eta^{(2)}|_{I_m}$$
$$\|\eta\|^2_{L^2} \leq 2(\|\eta^{(1)}\|^2_{L^2} + \|\eta^{(2)}\|^2_{L^2})$$

where we have [11]

$$\|\eta^-_m\|^2 \leq Ch^{2\mu}|u(t_m)|^2_{H^\mu(K)}$$
$$\int_{I_m} \|\eta^{(1)}\|^2 \leq C^2 h^{2\mu}|u|^2_{L^2(I_m;H^\mu(K))}$$
$$\int_{I_m} |\eta^{(1)}|^2_{H^1(K)} \leq C^2 h^{2(\mu-1)}|u|^2_{L^2(I_m;H^\mu(K))}$$
$$\int_{I_m} \|\eta^{(2)}\|^2 \leq C^2 \tau_m^{2(q+1)}|u|^2_{H^{q+1}(I_m;L^2(K))}$$
$$\int_{I_m} \|\eta^{(2)}\|^2_{H^1(K)} \leq C^2 \tau_m^{2(q+1)}|u|^2_{H^{q+1}(I_m;H^1(K))}. \tag{6.57}$$

The various terms containing $\eta$ are estimated as follows, defining $\bar{h} = \max_m h_m$

$$\sum_{m=1}^{M} \tau_m \|\eta^-_{\rho,m-1}\|^2 \leq C \sum_{m=1}^{M} \tau_m h_m^{2\mu}|u(t_{m-1})|^2_{H^\mu(K)} \leq C \sup_t |u(t)|^2_{H^\mu(K)} \bar{h}^{2\mu} \sum_{m=1}^{M} \tau_m$$
$$\leq CT\bar{h}^{2\mu}\|u\|^2_{C([0,T];H^\mu(K))}$$

and

$$\sum_{m=0}^{M} \|\eta^{-}_{\rho,m}\|^2 \leq \|\eta^{-}_{\rho,0}\|^2 + C\sum_{m=1}^{M} h_m^{2\mu}|u(t_m)|^2_{H^\mu(K)} \leq C\sup_t |u(t)|^2_{H^\mu(K)} \bar{h}^{2(\mu-1)} \sum_{m=1}^{M} h_m^2$$

If we have that

$$\frac{h_m^2}{\tau_m} \leq C \tag{6.58}$$

then

$$\sum_{m=1}^{M} \|\eta^{-}_{\rho,m}\|^2 \leq CT\bar{h}^{2(\mu-1)} \|u\|^2_{C([0,T];H^\mu(K))}.$$

Also for another term

$$\sum_{m=1}^{M} \int_{I_m} \|\eta_\rho\|^2_{H^1(K)} dt \leq 2\int_0^T \left(\|\eta_\rho^{(1)}\|^2_{H^1(K)} + \|\eta_\rho^{(2)}\|^2_{H^1(K)}\right) dt$$

and from (6.57) we have

$$\sum_{m=1}^{M} \int_{I_m} \|\eta_\rho\|^2_{H^1(K)} dt \leq C\Big(h^{2(\mu-1)}|u|^2_{L^2([0,T];H^\mu(K))} + \tau_m^{2(q+1)}|u|^2_{H^{q+1}([0,T];H^1(K))}\Big)$$

Based on the above analysis it can be seen the error is optimal in the norm defined in (6.55). We summarize the final error estimate as follows

**Theorem 6.4.2.** *Let* $\rho, q_x, q_y \in H^{q+1}(0,T;H^s(\Omega))$ *and* $\mu = min\{p+1, s\}$. *Then if the assumptions (i)-(iv) are satisfied and* $\tau$ *satisfy* (6.53) *and* (6.58) *and the local stabilization parameters are taken as* $\alpha = \mathcal{O}(h^{-1}), \beta = \mathcal{O}(h^{-1}), \beta' =$

$\mathcal{O}(h^{-1})$ *we obtain the following optimal error estimate*

$$
\begin{aligned}
&\|e^{-}_{\rho,M}\|^2 + \|e^{-}_{q_x,M}\|^2 + \|e^{-}_{q_y,M}\|^2 \\
&+ \sum_{m=1}^{M} \int_{I_m} \Big( \alpha \|(e_\rho - e_{\lambda_\rho})\|^2_{L^2(\Gamma_{h,n})} + \beta \|(e_{q_x} - e_{\lambda_{q_x}})\|^2_{L^2(\Gamma_{h,n})} \\
&+ \beta' \|(e_{q_y} - e_{\lambda_{q_y}})\|^2_{L^2(\Gamma_{h,n})} \Big) dt + \int_{I_m} (\|\xi_\rho\|^2 + \|\xi_{q_x}\|^2 + \|\xi_{q_y}\|^2) dt \\
&+ \sum_{m=1}^{M} \int_{I_m} \Big( \frac{c_1}{\epsilon_1} \|e_{\sigma_x}\|^2 + \frac{c_2}{\epsilon_2} \|e_{\sigma_y}\|^2 + c_3 \epsilon_1 \|\nabla e_{q_x}\|^2 + c_4 \epsilon_2 \|\nabla e_{q_y}\|^2 \Big) dt \\
&\leq C\big( h^{2(\mu-1)} + \tau^{2(q+1)} \big).
\end{aligned}
$$

# Chapter 7

# Conclusion

This dissertation was about the development and implementation of Space-Time Hybridized Discontinuous Galerkin method for 1D and 2D shallow water equations from both the theoretical and computational perspective. One of the main advantage of hybrid implicit method is that we can bypass many limitation which are needed to have numerical stability in the case of (explicit in time) DG methods. For example, no need for limiter in order to converge, choosing huge time steps (in order of hours) to run the simulation, no need to add bottom friction or ramp-up functions in tidal simulation. These are all advantages with respect to DG codes, however the trade-off is the need for good initial guess for convergence in the Newton method.

## 7.1 Accomplishments

• STHDG method was formulated for 1D and 2D shallow water equations.

• For 1D and 2D SWE, a 2D and 3D finite element code were implemented from scratch and parallelized. In 3D case two different types of cube and prism elements were implemented.

• The codes were tested with benchmark problems and very promising results were obtained. We also ran the code for a practical tidal-flow simulation. The code was able to bypass many limitations in DG code such as no need for ramp-up or bottom friction. The initial results is promising but regarding the convergence of Newton method, we need to have a good initial guess otherwise we would end up with lowering the time step or using a damped version of Newton method which essentially would increase the number of iterations needed.

• An *a priori* error estimate was proved for 2D shallow water equations with optimal rate of convergence in an appropriate norm. There are some literature on *a priori* error estimate but mostly they are limited to scalar equation and/or linear polynomial in time. We have extended the method first to the system of equations, with arbitrary degree of polynomial in time and hybridized mixed DG method.

## 7.2 Future work

The future work would mainly focus on these directions

• Implementing a multi-grid method so that we can find a good initial guess for Newton method resulting in choosing bigger time steps and (expected) less computational time.

• Utilizing the inherent capabilities of space-time methods, for example in case of moving meshes and also h/p adaptivity in time.

• Proving/modifying the method to the case of an entropy stable scheme such that it discretely satisfies the entropy inequality.

• Running the code for hurricane simulation and/or storm surge modeling and comparing the results with that of DG code.

# Bibliography

[1] V. Aizinger and C. Dawson. A discontinuous Galerkin method for two-dimensional flow and transport in shallow water. *Advances in Water Resources.*, 25:67–84, 2002.

[2] VR Ambati and O. Bokhove. Space-time discontinuous Galerkin finite element method for shallow water flows. ;. *J Comput Appl Math*, 204:452–462, 2007.

[3] D.N. Arnold and F. Brezzi. Mixed and non-conforming finite element methods: implementation, post-processing and error estimates. *Modl. Math. Anal. Numr.*, 19:7–35, 1985.

[4] E. Audusse, F. Bouchut, M. Bristeau, R. Klein, and B. Perthame. A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows. *SIAM Journal on Scientific Computing*, 25:2050–2065, 2004.

[5] T. J. Barth and D. C. Jespersen. The design and application of upwind schemes on unstructured meshes. *AIAA*, 0366, 1989.

[6] A. Bermudez and M.E. Vasquez. Upwind methods for hyperbolic conservation laws with source terms. *Comput. and Fluids*, 23:1049–1071, 1994.

[7] R. Botchorishvili, B. Perthame, and A. Vasseur. Equilibrium schemes for scalar conservation laws with stiff sources. *Mathematics of Computation*, 72:131–157, 2003.

[8] S. Brenner and R. Scott. *The Mathematical Theory of Finite Element Methods*. Springer, 2008.

[9] F. Brezzi, J. Douglas, and L.D. Marini. Two families of mixed finite elements for second order elliptic problems. *Numer. Math.*, 47:217–235, 1985.

[10] A. Canestrelli, M. Dumbser, A. Siviglia, and E. F. Toro. Well-balanced high-order centered schemes on unstructured meshes for shallow water equations with fixed and mobile bed. *Advances in Water Resources*, 33::291–303, 2010.

[11] J. Cesenek and M. Feistauer. Theory of the space-time discontinuous Galerkin method for nonstationary parabolic problems with nonlinear convection and diffusion. *SIAM J. NUMER. ANAL*, 50:1181–1206, 2012.

[12] A. Cesmelioglu, B. Cockburn, N. C. Nguyen, and J. Peraire. Analysis of HDG methods for Oseen equations. *Journal of Scientific Computing*, 55:392–431, 2013.

[13] B. Chabaud and B. Cockburn. Uniform-in-time superconvergence of HDG methods for the heat equation. *Math. Comp.*, 81:107–129, 2012.

[14] K. Chrysafinos and N. J. Walkington. Error estimates for the discontinuous Galerkin methods for parabolic equations. *SIAM J. NUMER. ANAL*, 44:349–366, 2006.

[15] B. Cockburn and B. Dong. An analysis of the minimal dissipation local discontinuous Galerkin method for convection-diffusion problems. *Journal of Scientific Computing*, 32:233–262, 2007.

[16] B. Cockburn, B. Dong, and J. Guzman. A superconvergent LDG-hybridizable Galerkin method for second-order elliptic problems. *Math. Comp.*, 77:1887–1916, 2008.

[17] B Cockburn, B Dong, J Guzman, M Restelli, and R Sacco. A hybridizable discontinuous Galerkin method for steady-state convection-diffusion-reaction problems. *SIAM J. Sci. Comput.*, 31(5):3827–3846, 2009.

[18] B. Cockburn, J. Gopalakrishnan, and R. Lazarov. Unified hybridization of discontinuous Galerkin, mixed and continuous Galerkin methods for second order elliptic problems. *SIAM J. Numer. Anal*, 47(2):1319–1365, 2009.

[19] B. Cockburn, J. Gopalakrishnan, N. C. Nguyen, J. Peraire, and F. Sayas. Analysis of HDG methods for stokes flow. *Math. Comp.*, 80:723–760, 2011.

[20] B. Cockburn and V. Quenneville-Blair. Uniform-in-time superconvergence of the HDG methods for the acoustic wave equation. *Math. Comp.*,

83:65–85, 2014.

[21] B. Cockburn and C.W. Shu. The Runge-Kutta Discontinuous Galerkin method for conservation laws v. *Journal of Computational Physics*, 141:199–224, 1998.

[22] J.F. Colombeau. *New Generalized Functions and Multiplication of Distributions*. North Holland, 1984.

[23] J. Cote and A. Staniforth. An accurate and efficient finite-element global model of the shallow-water equations. *Monthly Weather Review*, 118:2707–2717, 1990.

[24] O. Delestre and F. Marche. A numerical scheme for a viscous shallow water model with friction. *Sci Comput*, 48:41–51, 2011.

[25] L. Demkowicz and J. Gopalakrishnan. A class of discontinuous Petrov-Galerkin methods. part i: The transport equation. *Computer Methods in Applied Mechanics and Engineering*, 199:1558–1572, 2010.

[26] L. Demkowicz and J. Gopalakrishnan. A class of discontinuous Petrov-Galerkin methods. part ii: Optimal test functions. *Numerical Methods for Partial Differential Equations*, 27:70–105, 2011.

[27] L. Demkowicz, J. Gopalakrishnan, and A. H. Niemi. A class of discontinuous Petrov-Galerkin methods. part iii: Adaptivity. *Applied Numerical Mathematics*, 62:396–427, 2012.

[28] D. Dutykh, R. Poncet, and F. Dias. The VOLNA code for the numerical modelling of tsunami waves: generation, propagation and inundation. *Eur J Mech-B/Fluids*, 30:598–615, 2011.

[29] H. Egger and J. Schoberl. A hybrid mixed discontinuous Galerkin finite-element method for convection-diffusion problems. *IMA J Numer Anal*, 30 (4):1206–1234, 2010.

[30] T Ellis, L Demkowicz, and J. Chan. Locally conservative discontinuous Petrov-Galerkin finite elements for fluid problems. Technical report, Institute for Computational Engineering and Sciences, 2015.

[31] A. Ern, S. Piperno, and K. Djadel. A well-balanced Runge-Kutta discontinuous Galerkin method for the shallow-water equations with flooding and drying. *International Journal for Numerical Methods in Fluids Explore this journal*, 58:1–25, 2008.

[32] C. Eskilsson and S.J. Sherwin. Triangular spectral/hp discontinuous Galerkin method for modelling 2d shallow water equations. *Int J Numer Methods Fluids*, 45:605–23, 2004.

[33] L. C. Evans. *Partial Differential Equations*. American Mathematical Society, 2010.

[34] M. Feistauer, V. Kucera, K. Najzar, and J. Prokopova. Analysis of space-time discontinuous Galerkin method for nonlinear convection-diffusion problems. *Numerische Mathematik*, 117:251–288, 2011.

[35] T. Gallouet, J.M. Herard, and N. Seguin. Some approximate Godunov schemes to compute shallow-water equations with topography. *Computers and Fluids*, 32:479–513, 2003.

[36] F.X. Giraldo, J.S. Hesthaven, and T. Warburton. Nodal high-order discontinuous Galerkin methods for the spherical shallow water equations. *Journal of Computational Physics*, 181:499–525, 2002.

[37] L. Gosse. A well-balanced flux-vector splitting scheme designed for hyperbolic systems of conservation laws with source terms. *Comput. Math. Appl.*, 39:135–159, 2000.

[38] L. Gosse. A well-balanced scheme using non-conservative products designed for hyperbolic systems of conservation laws with source terms. *Math. Models Methods Appl. Sci.*, 11:339–365, 2001.

[39] L. Gosse and A.Y. LeRoux. A well-balanced scheme designed for inhomogeneous scalar conservation laws. *C.R. Acad. Sci. Paris S er.I Math*, 323:543–546, 1996.

[40] J.M. Greenberg and A.Y. LeRoux. A well balanced scheme for the numerical processing of source terms in hyperbolic equations. *SIAM J. Numer. Anal.*, 33:1–168, 1996.

[41] R. Heikes and D. A. Randall. Numerical integration of the shallow-water equations on a twisted icosahedral grid.2. a detailed description

of the grid and an analysis of numerical accuracy. *Mon. Weather Rev*, 123:1881–1887, 1995.

[42] H. Hoteit, Ph. Ackerer, R. Mos, J. Erhel, and B. Philippe. New two-dimensional slope limiters for discontinuous Galerkin methods on arbitrary meshes. *Journal of Computational Physics*, 61:2566–2593, 2004.

[43] R. Jakob-Chien, J. J. Hack, and D. L. Williamson. Spectral transform solutions to the shallow water test set. *Journal of Computational Physics*, 119:164–187, 1995.

[44] G. Jiang and C.W. Shu. Efficient implementation of Weighted ENO schemes. *J Comput Phys*, 126:202–228, 1996.

[45] S. Jin. A steady-state capturing method for hyperbolic systems with geometrical source terms. *M2AN Math. Model. Numer. Anal.*, 35:631–645, 2001.

[46] L. Krivodonova, J. Xin, J.-F. Remacle, N. Chevaugeon, and J.E. Flaherty. Shock detection and limiting with discontinuous Galerkin methods for hyperbolic conservation laws. *Applied Numerical Mathematics*, 48:323–338, 2004.

[47] EJ Kubatko, JJ Westerink, and C Dawson. hp discontinuous Galerkin methods for advection dominated problems in shallow water flow. *Comput Methods Appl Mech Eng*, 196, 2006.

[48] R.J. Leveque. Balancing source terms and flux gradients in high-resolution Godunov methods: the quasi-steady wave-propagation algorithm. *J. Comput. Phys.*, 146:346–365, 1998.

[49] S-J Lin and B. Rood. An explicit flux-form semi-Lagrangian shallow water model on the sphere. *Quart. J. Roy. Meteor. Soc*, 123:2531–2533, 1997.

[50] R. Liska and B. Wendroff. Two-dimensional shallow water equations by composite schemes. *International journal for numerical methods in engineering*, 30::461–479, 1999.

[51] G. Dal Maso, P. G. Lefloch, and F. Murat. Definition and weak stability of nonconservative products. *Journal de mathmatiques pures et appliques*, 74:483–548, 1995.

[52] R. D. Nair, S. J. Thomas, and R. D. Loft. A discontinuous Galerkin global shallow water model. *Monthly Weather Review*, 133:814–828, 2005.

[53] N. C. Nguyen, J. Peraire, and B. Cockburn. An implicit high-order hybridizable discontinuous Galerkin method for linear convection-diffusion equations. *Journal of Computational Physics*, 228:3232–3254, 2009.

[54] N. C. Nguyen, J. Peraire, and B. Cockburn. An implicit high-order hybridizable discontinuous Galerkin method for nonlinear convection-diffusion equations. *Journal of Computational Physics*, 228:8841–8855, 2009.

[55] J. Nitsche. ber ein variationsprinzip zur lsung von dirichlet-problemen bei verwendung von teilrumen, die keinen randbedingungen unterworfen sind. *Abhandlungen aus dem Mathematischen Seminar der Universitt Hamburg*, 36:9–15, 1971.

[56] T.H.H. Pian. Derivation of element stiffness matrices by assumed stress distributions. *AIAA*, 2:1333–1336, 1964.

[57] P.A. Raviart and J.M. Thomas. A mixed finite element method for second order elliptic problems. *Mathematical Aspects of Finite Element Methods*, 606:292–315, 1977.

[58] P.A. Raviart and J.M. Thomas. Primal hybrid finite element methods for 2nd order elliptic equations. *J. Math. Comput.*, 31:391–413, 1977.

[59] S. Rhebergen, B. Cockburn, and J. J. W. Van Der Vegt. A space-time discontinuous Galerkin method for the incompressible Navier-Stokes equations. *Journal of Computational Physics*, 233:339–358, 2013.

[60] C. Ronchi, R. Iacono, and P. S. Paolucci. Finite difference approximation to the shallow water equations on a quasi-uniform spherical grid. *High-Performance Computing and Networking*, 919:741–747, 2006.

[61] D. Schwanenberg and M. Harms. Discontinuous Galerkin finite-element method for transcritical two-dimensional shallow water flows. *J Hydraul Eng*, 130:412–21, 2004.

[62] S.C. Soon, B. Cockburn, and H. K Stolarski. A hybridizable discontinuous Galerkin method for linear elasticity. *International journal for numerical methods in engineering*, 80:1058–1092, 2009.

[63] H. Tang. Solution of the shallow-water equations using an adaptive moving mesh method. *Int. J. Numer. Meth. Fluids*, 44:789–810, 2004.

[64] A. Vasseur and Ch. Yu. Existence of global weak solutions for 3D degenerate compressible Navier-Stokes equations. *Inventiones mathematicae*, pages 1–40, 2016.

[65] B. Fraejis De Veubeke. Displacement and equilibrium models in the finite element method. *International journal for numerical methods in engineering*, 52:287–342, 2001.

[66] Y. Xing and X. Zhang. Positivity-preserving well-balanced discontinuous galerkin methods for the shallow water equations on unstructured triangular meshes. *Journal of Scientific Computing*, 57:19–41, 2013.

[67] Ya. B. Zeldovich and Yu. P. Raizer. *Physics of Shock Waves and High-Temperature Hydrodynamic Phenomena*. Dover, 2002.

[68] J. Zitelli, I. Mugaa, L. Demkowicz, J. Gopalakrishnan, D. Pardo, and V.M. Calo. A class of discontinuous Petrov-Galerkin methods. part iv: The optimal test norm and time-harmonic wave propagation in 1d. *Journal of Computational Physics*, 230:2406–2432, 2011.

[69] C. Zoppou and S. Roberts. Numerical solution of the two-dimensional unsteady dam break. *Applied Mathematical Modelling*, 24:457–475, 2000.